ESD-TR-67-202

PAPERS ON AUTOMATIC LANGUAGE PROCESSING

SELECTED COLLECTION STATISTICS AND DATA ANALYSES

Paul E. Jones
Vincent E. Giuliano
Robert M. Curtice

February 1967

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Mass.

(Prepared under Contract No. AF 19(628)-3311 by Arthur D. Little, Incorporated, Cambridge, Mass.)

LEGAL NOTICE

When U.S. Government drawings, specifications or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

OTHER NOTICES

Do not return this copy. Retain or destroy.

ESD-TR-67-202, Vol. I

PAPERS ON AUTOMATIC LANGUAGE PROCESSING

SELECTED COLLECTION STATISTICS AND DATA ANALYSES

Paul E. Jones
Vincent E. Giuliano
Robert M. Curtice

February 1967

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts

(Prepared under Contract No. AF 19(628)-3311 by Arthur D. Little, Inc., Cambridge, Massachusetts)

FOREWORD

This is Volume I of a group of three containing work on automatic language processing. This work involved both theoretical and experimental investigations of natural language characteristics as well as the properties of actual and laboratory-modified document collections and retrieval situations. Titles of the volumes are:

Volume I	Selected Collection Statistics and Data Analyses
Volume II	Linear Models for Associative Retrieval
Volume III	Development of String Indexing Techniques

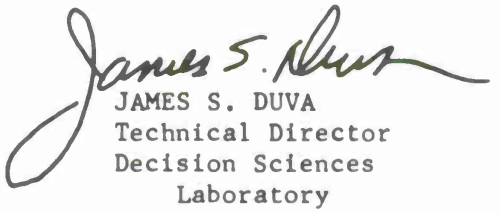
This work was conducted in support of Project 2806, Task 280601 by Arthur D. Little, Inc., 35 Acorn Park, Cambridge, Massachusetts under Contract AF 19 (628) - 3311. Our internal code for this contract is C - 65850. The work was also supported in part under Contract AF 19 (628) - 4067.

The program was monitored for the U. S. Air Force by John B. Goodenough ESVPD and was principally performed during the period January 1965 to December 1966, and the draft report was submitted on 15 January 1967. We wish to acknowledge the important contribution of Philip Hankins, Inc. in performing part of the computer programming under subcontract to us.

The cooperation between the Science and Technology Information Division of NASA and the Decision Sciences Laboratory has been of immense value to the research reported here. In particular programs developed under our Contract NASW-1051 with NASA have been used in some of the investigations reported here, and conversely.

This Technical Report has been reviewed and approved.

WALTER E. ORGANIST
Project Officer
Decision Sciences Laboratory



JAMES S. DUVA
Technical Director
Decision Sciences
Laboratory

ABSTRACT

As part of a research program aimed at determining the parameters influencing the effectiveness of a message retrieval system, a collection of 10,000 technical abstracts was indexed and retrieval experiments were conducted with them. Since part of the work involved the development and test operation of an associative retrieval system, basic data about the distribution of words and word strings were gathered in preparing the system for test and trial. These statistics were thought to be of possible interest to other workers in the field and are gathered as a series of loosely connected papers in this folume under the following groupings: Characteristics and Indexing of GE Data Base; Comparison of Manual and Machine Selected Vocabularies; Vocabulary Distribution Studies, and Studies of Content Bearing Units in Text.

PREFACE

In the course of five years of work on automatic language processing which involved both theoretical and experimental investigations of natural language characteristics and the properties of actual and laboratory-modified document collections and retrieval situations, we have written a number of technical papers, working papers, and internal notes whose possibly useful content has by no means been completely revealed in the course of publishing a series of Project reports.

We feel it desirable to make this material available to other workers in the field partly to discharge the normal obligations to publish and partly with the hope of stimulating further work in a new, difficult, and potentially very rewarding field. We do not, however, feel obligated to impose an artificial structure on this work which is essentially supportive in nature. Therefore, this series of volumes is a collection of papers, loosely grouped into areas, but not otherwise intended to demonstrate coherence, some of which are in support of, and others peripheral to our published reports.

This is Volume I of a group of three whose titles are:

VOLUME I	Selected Collection Statistics and Data Analyses
VOLUME II	Linear Models for Associative Retrieval
VOLUME III	Development of String Indexing Techniques

TABLE OF CONTENTS

SELECTED COLLECTION STATISTICS AND DATA ANALYSES

INTRODUCTION

As part of a research program aimed at determining the parameters influencing the effectiveness of a message retrieval system we automatically indexed and conducted retrieval experiments on a collection of 10,000 technical abstracts. Each of these abstracts was regarded as an assertive message in its own right -- i.e., an information bearing unit not necessarily related to other messages in the set.

Part of our work involved the development and test operation of an associative retrieval system to operate on this collection. This system, in the form in which it was tested, responded to full text English queries by ranking the 10,000 stored messages according to the relevance of each to the submitted request.

In preparing the system for test and trial, basic data about the distribution of words and word strings were gathered. Some of the statistics were obtained because they were needed for decisions we made along the way; other statistics were gathered largely because it was natural or easy to obtain them as a byproduct of the processing.

The tapes already employed in a large operational coordinate retrieval system were purchased for our experimental use. This collection was chosen in part because it had the following desirable attributes:

a. It was developed independently of Arthur D. Little, Inc., and reflects a real information retrieval system in current use.

b. The collection size (c. 70,000 documents) is sufficiently large to reflect a "mature" retrieval system.

c. It is a "pure" coordinate system in the sense that no hierarchical indexing strategy is used.

d. Those responsible for the system have resisted attempts to use terms that are not single words.

e. The vocabulary of term-usage is relatively open-ended, with many synonyms being admissible, and the rank-frequency characteristic of term usage tends to behave according to Zipf's law, like natural language. The vocabulary therefore is of the general kind which can be obtained using automatic indexing techniques.

f. Abstracts of the majority of the documents indexed in the collection are available in machine-readable form.

Two magnetic tapes were obtained. The first of these is an index tape which lists, for each of some 70,000 documents, the index terms that were assigned to them. The second set of tapes consists of English abstracts of the information contained in about 45,000* of the documents. 4,500-5,000 index terms comprised the vocabulary of terms used to index the documents.

The association programs for the IBM 7090 existing at the time could not handle a collection larger than about 10,000 documents indexed by about 1,000 terms. Accordingly, it was necessary to extract a portion of the given data to serve as input data for our experiments.

Two subcollections were derived and are referred to throughout this Volume. They are:

GE 1-A Indexing Vocabulary - The collection of 70,000 G.E. documents was manually indexed from a vocabulary of 4826 Uniterms. This vocabulary was partitioned at ADL, and we identified a group of about 1560 primarily metallurgical Uniterms we wished to exclude from consideration. By choosing essentially every third term from the group of 3266 remaining terms, a 1087 term sample of the "interesting" terms was obtained. This sample is GE 1-A.

GE 2-A Indexing Vocabulary - GE 2-A is the set of 999 vocabulary items used for automatic indexing of the collection of GE abstracts. Whereas the Uniterms which form the basis of GE 1-A were assigned by human indexers on the basis of reading the whole document, the GE 2-A terms are the 999 highest-frequency content words which appear in the texts of 45,000 G.E. abstracts. Singular and plural forms were coalesced.

This volume gathers together some of the more interesting statistical observations, frequency data and side effects that were obtained during our work with this 446,097 word corpus of technical text. Section I describes the data, the processing, and some frequency distribution studies. Section II is a paper which compares the manually-selected vocabulary and one selected by automatic processing. Section III is a group of papers dealing with vocabulary distribution studies including token frequencies and entropy calculations, and the fitting of the Herdan-Waring distribution and Zipf curves for the vocabulary. In Section IV the papers deal with studies of content bearing units in Text. A partial listing of the data is given and the units which were discovered as content bearing were contrasted with two word index terms in the NASA Vocabulary.

* All abstracts in the collection were provided except those which are under security classification or considered proprietary by the company which provided the data.

SECTION I

CHARACTERISTICS AND INDEXING OF GE DATA BASE

Word Units, Frequency Counts, and Machine Indexing of GE 2 Data Base*

A. INTRODUCTION

The present note describes and documents the computer processing which was performed, during the second half of 1964, as an adjunct to research on techniques for automatic message retrieval. During this period, as described in detail below, it was found desirable to investigate procedures for automatically indexing short messages by selecting words (or strings of words) from the text of the message to serve as index terms. The "messages" available in sufficient numbers on magnetic tape consisted of the abstracts of technical articles. This note describes the processing of these abstracts for the purpose of obtaining:

1. Detailed frequency data about the recurrence patterns of word strings. (See TN CACL-10 and Volume III this series.)

2. An operational retrieval system based on automatically indexing these abstracts. (See TN CACL-11.)

The data base chosen consisted of a subset of about 10,000 abstracts chosen from the G.E. collection (See TN CACL-12) of 45,000 abstracts dealing with topics about design, construction and testing of aerospace vehicles. This data will be referred to as the G.E. 2 data base. These abstracts of technical articles can be viewed as messages, for they report in compact and precise form a piece of factual information, namely the content of the document of which they are an abstract.

Experiments on automatically recognizing strings of words as conceptual units, based on knowledge only of the frequencies of the strings and their substrings, had previously been tried on a smaller text with promising results** It was therefore desirable to ascertain whether these techniques were of significant value when applied to a data base sufficiently large to yield conclusive results. The procedures for obtaining the desired frequency data for recurrent word-strings of length up to 4 words are described in this note. The investigations for which these data are used are discussed in TN CACL-4, Volume III, this series.

---

*Issued on March 9, 1965 to a limited distribution by Joyce S. Mehring as Technical Note CACL-13.

** This work is described in TN CACL-4, Volume III, this series.

Section I: CACL-13

The production of an operational retrieval system based on automatic indexing was desired as part of research on techniques for evaluating retrieval systems. The orientation of this research made it desirable to prepare two operating systems in order that their performance could subsequently be compared with each other and with the performance of human beings who would also "conduct retrieval" on the same collection. This line of research is discussed in TN CACL-11. Both a completely operational associative retrieval system and the coordinate system which appears as a by-product were needed for this work. This section also describes the procedures whereby single words were used to index the collection automatically for these ends.

B. COMPUTER PROCESSING OF G.E. 2 DATA

The description of computer processing of the G. E. 2 Data is discussed in three sections. Section describes the selection of a subset of the G. E. abstracts to be called the G.E. 2 data base and the generation of all four-word strings appearing in this subset. Section describes the procedures for generating distinct four-word, three-word, two-word and one-word strings, and the frequencies of strings and ordered substrings. The procedures used to select a set of single words to be used as index terms and the procedure for assigning index terms to documents are discussed in Section .

1. Selecting G. E. 2 Abstracts and Generating Four-Word Strings

To sample the G. E. abstracts and generate four-word strings, a 7090 computer program, CNTXT, was designed which operated on tapes containing the G. E. abstracts and a tape containing exception words and carried out the following procedures:

selected from a G. E. subcollection of 45,000 abstracts every fourth abstract containing more than six lines and recorded the 10,289 selected abstracts on magnetic tape;

for each selected abstract, produced the G. E. abstract number and a sequential number on magnetic tape;

for each selected abstract, produced on magnetic tape four-word strings with the G. E. abstract number in which the string appeared. All words which were exception words were marked with a leading blank. Approximately 446,000 four-word strings were produced.

a. Definition of Four-word String

A four-word string or four-word context consists of four contiguous words within an abstract. Punctuation marks are not considered to be words; hence two words are contiguous

even if separated by punctuation marks. For every word in the abstract, except words in the first line, a four-word string is produced. Dummy strings are produced for the last three words in an abstract by using blanks as the terminal words in the four-word string. The text of the abstract, excluding the first line, is scanned character by character and "words: are recognized. In general these "words" correspond to words as usually recognized, however a precise definition of "word" in this exercise is arrived at by applying the definitions and rules given below.

(1) Method of Word Recognition

Definitions

S - Special characters [' - .) , blank # / (]

A - All other characters appearing in the text

E - Special signals not appearing in the text

$[join_1, join_2]$

Classification of characters for the Word Recognition Procedure

Following is a list of characters and their classifications:

Type 1	' -
Type 2	.
Type 3	),
Type 4	blank #
Type 5	/ (
Type 6	All characters in A
Type 7	$join_1$
Type 8	$join_2$

"Reduction" Rules

Certain pairs of characters from S and E when appearing as contiguous characters in the text are classified as a single character from S or E. The "reduction" rules used in these classifications are stated below.

Type 1, Type 4 $\longrightarrow$ Type 7
Type 7, Type 4 $\longrightarrow$ Type 7

Type 4, Type 1 ⟶ Type 8
Type 2, Type 4 ⟶ Type 4
Type 3, Type 4 ⟶ Type 4
Type 4, Type 4 ⟶ Type 4
Type 5, Type 4 ⟶ Type 4
Type 8, Type 4 ⟶ Type 4

Special Character Configurations for the Word Recognition Procedure

Break configuration: a Type 6 character preceded by a Type 1, 3, 4, or 5 character.

Concatenation configuration: a Type 6 character preceded by a Type 7 or 8 character.

Decimal configuration: a Type 6 character preceded by a Type 2 character.

(2) Word Recognition Procedure

Each character in the text excl ding the first line of each abstract is examined in a left to right scan. As each character is scanned the classification of the preceding character is available. If the character under examination is a character from the set S, a reduction rule is applied if appropriate and the resultant classification is retained, otherwise the classification of the special character is retained. If a character in A is examined, that is, a Type 6 character, it is appended to the current character string and its classification is retained unless

(a) A break configuration has been encountered in which case the present character string is said to be a word and a new character string is started by the Type 6 character in the configuration.

(b) A concatenation configuration is encountered in which case the Type 6 character in the configuration is appended to the present character string.

(c) A decimal configuration is encountered in which case the decimal point and Type 6 character are appended to the character string. In all cases the scan is continued.

After each word is recognized it is examined to see if it is one of 240 exception words. If the word is an exception word, it is marked by a leading blank. Each

word produced consists of a maximum of 12 characters. If a word recognized in the text consists of more than 12 characters, only the first 12 characters are used in producing a word to be used in a word string.

2. Generating Contexts with Frequency Information

The next objective was to produce four-word, three-word, two-word and one-word contexts, each with information about the frequency of occurrence of the contexts and substrings contained in the contexts. Let A, B, C, S represent four words and let

f_A represent the number of times the word A occurs in the text.

f_{AB} represent the number of times the contiguous pair AB occurs in the text, etc.

Then the content of the four desired lists can be summarized in the following way:

The one-word context list contains for each distinct word:

A f_A

The two-word context list contains for each distinct contiguous pair:

A B f_A f_B f_{AB}

The three-word context list contains for each distinct contiguous triplet:

A B C f_A f_B f_C f_{AB} f_{BC} f_{ABC}

The four-word context list contains for each distinct contiguous quadruplet:

A B C D f_A f_B f_C f_D f_{AB} f_{BC} f_{CD} f_{ABC} f_{BCD} f_{ABCD}

a. Production of Distinct Strings

To obtain the context lists described above, the first major step was to obtain distinct four-word, three-word, two-word and one-word strings with their frequency of occurrence. A 7090 computer program, SQUISH, was designed to operate on alphabetically ordered four-word strings to produce the four sets of distinct strings and frequencies. In preparation for SQUISH, the tape of four-word strings produced by CNTXT was sorted into alphabetical order on the strings by the SORT program in the IBM Basic Monitor System IBSYS. This sorted tape was then used by SQUISH to produce four lists on magnetic

tape. The content of the four lists (four tapes) is described below:

Distinct one-word strings and frequency of occurrence:
A f_A

Distinct two-word strings and frequency of occurrence:
A B f_{AB}

Distinct three-word strings and frequency of occurrence:
A B C f_{ABC}

Distinct four-word strings and frequency of occurrence:
A B C D f_{ABCD}

To be placed on a list a string had to occur a minimum number of times. These minimum frequencies were given as input data to SQUISH. The minimum frequencies used and the resulting number of strings produced are summarized below. The number of strings is approximate.

23,600 distinct one-word strings with $f_A \gg 1$

48,000 distinct two-word strings with $f_{AB} \gg 2$

11,700 distinct three-word strings with $f_{ABC} \gg 3$

3, 350 distinct four-word strings with $f_{ABCD} \gg 3$

b. Production of Final Context Tapes

The list of one-word strings produced by SQUISH is in the form desired for the one-word context list and hence the tape containing that list is the one-word context tape. The three remaining context lists could be produced from the information available on the tapes produced by SQUISH. The procedure for producing these desired lists was to obtain the frequency of substrings of a K-word string from available frequency information about K-1 word strings. The frequencies of all substrings of a string were available since in this case the minimum frequency of a K-word string is greater than or equal to the minimum frequency of a K-1 word string. As pointed out in the previous section

$$\min\left\{f_{ABCD}\right\} = 3 = \min\left\{f_{ABC}\right\} = 3 > \min\left\{f_{AB}\right\} = 2 > \min\left\{f_A\right\} = 1.$$

(1) First Merge Phase

Two 7090/94 Computer programs were designed to do "merging". The first "merging" operations were carried out by a 7090 computer program MERGE I which produced from the four tapes generated by SQUISH three tapes designated here by T2', T3' and T4' whose content can be summarized in the following way:

T2' Two-word string tape A B f_A f_{AB}

T3' Three-word string tape A C C f_A f_{AB} f_{ABC}

T4' Four-word string tape A B C D f_A f_{AB} f_{ABC} f_{ABCD}

(2) Second Merge Phase

The second set of merging operations was carried out by a 7090 program MERGEG. This program produced from a specially ordered K-word string tape and a K-1 word final context tape a K word string tape with all the desired frequency information. This newly generated tape when sorted into alphabetical order on entire word strings was the final K-word context tape.

As a preliminary step of data preparation for MERGEG the tapes produced by MERGE I were sorted in the following way:

T2' was sorted into alphabetical order on the second word.

T3' was sorted into alphabetical order on the two terminal words.

T4' was sorted into alphabetical order on the three terminal words.

The IBSYS SORT program was used. The resultant tapes will be designated here as ST2', ST3' and ST4'.

(3) Final Two-Word, Three-Word and Four-Word Context Tapes

MERGEG and the IBSYS SORT programs were then used alternately to produce the desired lists. The final one-word context tape and ST2' were input to MERGEG which produced a tape containing for every entry:

A B f_A f_B f_{AB}.

This output tape, ordered like ST2', when sorted into alphabetical order on the entire word string was the final two-word context tape.* MERGEG then operated on the final two-word context tape and ST3' to produce a tape which contained for every entry:

$$A\ B\ C\ f_A\ f_B\ f_C\ f_{AB}\ f_{BC}\ f_{ABC}.$$

This tape, ordered like ST3', when sorted into alphabetical order on the entire word string was the final three-word context tape.* The final three-word context tape and ST4' were used by MERGEG to produce a tape which contained for every entry:

$$A\ B\ C\ D\ f_A\ f_B\ f_C\ f_D\ f_{AB}\ f_{BC}\ f_{CD}\ f_{ABC}\ f_{BCD}\ f_{ABCD}.$$

This tape, ordered like ST4", when sorted into alphabetical order on the entire word string was the final four-word context tape.

3. Indexing Documents

To produce a collection of abstracts automatically indexed by single words selected from text, frequency information was used for selecting index terms and a computer program was disegned to assign the resulting index terms to the abstracts in which they appeared.

a. Selection of Index Terms

The index terms were selected from among the one-word contexts by the following criteria:

The singular and plural forms of words occurred jointly a total of 56 or more times in the text;

The first character of the word was a letter A through Z. This latter constraint eliminated exception words and numbers from the set of index terms.

Using these criteria, 1434 distinct words were selected, counting both singular and plural forms. After coalescing singular and plural forms, that is, assigning the same representative form to both the singular and plural form of a word, there were 999 representative index terms. If a word occurring in the list of 1434 terms was found in an abstract, the representative term in the list of 999 terms was assigned as an index term. The set of index terms each with its representative was recorded on a dictionary tape.

b. Indexing Procedure

A 7090 computer program INDEX was designed which assigned an index term to an abstract if the term appeared in the ab-

* The number of entries on these tapes is the same as the number on the corresponding 1, 2, 3, and 4-word tapes produced as the output of SQUISH.

stract. From the four-word string tape produced by CNTXT and a tape of dictionary terms selected according to the above criteria, the program produced a tape containing abstract number-term number pairs. This tape contained about 222,000 pairs, that is, about 220,000 words in the text were index terms. This tape was then sorted into abstract number term number order and duplicate pairs were eliminated by a 7090/94 computer program ELIM. Hence if an index term appeared more than once in an abstract, only one pair entry was retained. This pair tape was then used by an existing program to produce a Packed Document Term Matrix. This Packed Document Term Matrix was the input to a set of existing programs which produced a document-term tape and a word-association matrix for use in retrieval. Among the first steps in the production of the association matrix, index-word pairs were generated based on co-occurrence of index terms within an abstract. About 2,933, 000 index-word pairs were generated for this data base.

FUNCTION WORDS*

The following words comprised the exception list of function words in the processing which was described in Technical Note CACL-13. We are indebted to H. Rubenstein for this list.

A. Alphabetic Listing of Function Words

ABOUT
ABOVE
ACROSS
AFTER
AGAINST
ALL
ALMOST
ALONE
ALONG
ALSO
ALTHOUGH
ALWAYS
AMONG
AM
AND
ANOTHER
AN
ANYBODY
ANYONE
ANY
ANYTHING
ANYWHERE
APART
ARE
AROUND
A
ASIDE
AS
AT
AWAY
AWFULLY
BECAUSE
BEEN
BEFORE
BEHIND
BEING
BELOW
BE
BETWEEN
BEYOND
BOTH
BUT
BY
CANNOT
CAN
COULD
DID
DOES
DOING
DONE
DO
DOWN
DURING
EACH
EITHER
ELSE
ELSEWHERE
ENOUGH
ETC
EVEN
EVER
EVERYONE
EVERY
EVERYTHING
EVERYWHERE
EXCEPT
FEW
FOR
FORTH
FROM
FURTHERMORE
GET
GETS
GOT
HAD
HARDLY
HAS
HAVE
HAVING
HENCE
HEREIN
HERE
HER
HERSELF
HE
HIM
HIMSELF
HIS
HITHER
HOWBEIT
HOWEVER
HOW
IF
INASMUCH
INDEED
INNER
IN
INSOFAR
INSTEAD
INTO
INWARD
I
IS
IT
ITSELF
ITS
JUST
KEEP
KEPT
LEAST
LESS

* Issued on April 15, 1965 to a limited distribution by Paul E. Jones as a Supplement to Technical Note CACL-13

LEST
MANY
MAY
ME
MIGHT
MINE
MOREOVER
MORE
MOST
MUCH
MUST
MY
MYSELF
NEITHER
NEVERTHELES
NEXT
NOBODY
NONE
NOR
NO
NOTHING
NOT
NOWHERE
OF
OH
ONE
ONES
ONLY
ON
ONTO
OR
OTHER
OTHERS
OTHERWISE
OUGHT
OUR
OURSELVES
OURS
OUTSIDE
OVER
OWN
PER
PLEASE
PLUS
QUITE
RATHER
REALLY
RIGHT
SELF
SELVES
SEVERAL
SHALL
SHE
SHOULD
SINCE
SIX
SOMEBODY
SOME
SOMETHING
SOMETIMES
SOMEWHAT
SO
STILL
SUCH
TEN
THAN
THAT
THEIR
THEIRS
THEM
THEMSELVES
THENCE
THEN
THEREBY
THEREFORE
THERE
THE
THESE
THEY
THIS
THOSE
THOUGH
THROUGHOUT
THUS
TOGETHER
TOO
TO
TOWARD
TWO
UNDERNEATH
UNDER
UNLESS
UNTIL
UNTO
UPON
UP
UPWARD
US
VERY
WAS
WELL
WERE
WE
WHATEVER
WHAT
WHENCE
WHENEVER
WHEN
WHERE
WHEREVER
WHETHER
WHICH
WHILE
WHOM
WHO
WHOSE
WHY
WILL
WITHIN
WITHOUT
WITH
WOULD
YES
YET
YOUR
YOURSELF
YOURSELVES
YOURS
YOU

B. Frequency of Function Words in Corpus G.E.-2

A partial list of the frequencies of occurrence in this text of the most frequent function words follows.

OF	30170	THIS	374
AND	16622	NO	372
THE	8911	UP	356
IN	8879	WERE	355
TO	8045	THAN	353
FOR	7406	ITS	347
A	6232	OTHER	331
ON	4634	WHEN	303
WITH	3844	HAS	292
AT	2770	DURING	286
BY	2759	HAVE	282
FROM	1807	BOTH	272
IS	1704	I	269
AS	1689	THESE	268
AN	1474	ALL	260
WHICH	1180	IT	260
BE	1077	SEVERAL	260
ARE	996	MAY	255
THAT	954	BEEN	250
OR	906	ALSO	241
TWO	899	ONLY	241
BETWEEN	676	THEIR	238
UNDER	596	MORE	225
NOT	558	PER	225
SOME	545	SUCH	222
CAN	479	BUT	212
OVER	424	ABOUT	204
INTO	410	HAVING	203
WAS	398	WITHOUT	184
ONE	381	UPON	175

(Count = 60)

Section I: Supplement 1 to CACL-13

The function words occurring between 12 and 1 times can be listed in frequency order.

ALWAYS	12	NOTHING	4
WE	12	ONTO	4
GET	11	SOMETIMES	4
QUITE	11	YOU	4
FORTH	9	APART	3
INSTEAD	9	HENCE	3
JUST	9	NEITHER	3
NOR	9	UNLESS	3
EVERY	8	YOU	3
HEREIN	8	DOING	2
WHY	8	ELSEWHERE	2
YET	8	EVERYWHERE	2
AWAY	7	MINE	2
OUR	7	OH	2
THEREBY	7	THEMSELVES	2
WHO	7	UPWARD	2
NONE	6	WHATEVER	2
ONES	6	ANYTHING	1
SOMEWHAT	6	ASIDE	1
THOUGH	6	GETS	1
EVER	5	HER	1
INWARD	5	HIM	1
KEEP	5	INSOFAR	1
KEPT	5	LEST	1
OTHERWISE	5	SOMETHING	1
THEREFORE	5	WHEREVER	1
ME	4		

(Count = 53)

Section I: Supplement 2 to CACL-13

ESTIMATING RECURRENCE OF MISSPELLINGS IN CORPUS G.E.-2 *

A. Anomalies among Words with Frequency 2

The following "words" were considered anomalous (by me) among the types which occurred twice in Corpus G.E.-2.

ABL	
AB	
ACIER	
ADVANM	ADVANCE ?
AGCL	
AGIG	
ALCHOL	ALCOHOL ?
ALCUMG	
ALLM	ALUM ?
ALUN	ALUM ?
ALYER	LAYER ?
ANALYSI	ANALYSIS ? or Russian Word ?
ANALYT	
ANAM	
ANF	
ANOCUT	
APPROXM	APPROX. ?
APRALLEL	PARALLEL ?
APUS	
ARO	
ARTIFICAL	ARTIFICIAL ?
ASCAST	
AT320	AT 320 ?
AT423F	AT 423 F ?
ATA	AT A ?
ATED	-ATED ?
AVON	
AVOVE	ABOVE ?
AWS	
BARIA	
BEHAVIOUS	BEHAVIOUR ?
BOLTZMAN	BOLTZMANN ?

* Issued on April 16, 1965 to a limited distribution by Paul E. Jones as a Supplement to Technical Note CACL-13

BOURDON	
BOVERI	
BRITTEL	BRITTLE ?
CALCULATIONO	CALCULATION OF ?
CALIBRAM	
CAPACITIVELY	?
CAPACITIVE	?
CARACTERISTIC	CHARACTERISTIC ?
CASCADEM	CASCADED ?
CAVIATATION	CAVITATION ?
CECOSTAMP	?
CENM	CENT ?
CENTRATION	
CHARACTERM	CHARACTER ?
CHARACTERTIC	CHARACTERISTIC ?
CIRM	
C.D	C.D. ?
C.W	C.W. ?
CLOSEDM	CLOSED ?
CLUDING	INCLUDING ?
COATINS	COATINGS ?
COEFFICEINTS	COEFFICIENTS ?
COMBUTION	COMBUSTION ?
COMME	
COMMINUTED	
COMPLEETE	
COMPLES	
COMPRESSEUR	French ?
COMPRESSORE	Italian ?
COMPTOIR	
COMPUM	
CONCN	
CONDI	CONDITION ?
CONDITIO	"
CONDITI	"
CONDITONS	CONDITIONS ?
COND	CONDITION ?

Summary, 450 words checked, 71 probable misspellings. (16 per 100)

B. Anomalies Among Words with Frequency 3

AEDC	
ATIONS	-ATIONS ?
ATURE	-ATURE ?
AUBES	

B
BRER
BULENT
BURING
BUTION
CERTAINS
CHAUD
CLUDED
COEFFICIEN
COMMERICAL
COMPLICANCE
COMPUTOR
COMUSTION
CONDIM
CONSIDM
CONSTIM
DEFORMAM
DESM
DETERMING

300 checked, 22 probable errors (7.3 per 100)

C. Anomalies Among Words Occurring 5 Times

APPENDIXES
ARY
BLASIUS
BUNA
DIFFERM
ENM
ENTRE
FORMANCE
ISTICS
LATED

300 tested, 9 anomalies (3 per 100)

D. Anomalies Among Words of Frequency 8

ATION	-ATION ?
CD	
FIZ	
MENTS	-MENTS ?
MISSLE	MISSILE ?
PREM	
SNECMA	

THERMOM
TRANSM

300 tested, 9 anomalies (3 per 100)

E. Anomalies Among Words of Frequency 10

LETUDE
NIMONIC
ONERA
PERM

~300 tested frequency 10 plus above, 4 anomalies

F. Estimating the Error Rate in Corpus G.E.-2

It would be useful to have an estimate of the number of misspelled words per 100 words of running text. An estimate of this can be obtained very crudely using the data so far acquired if they are plotted as in the attached figure.

The plot shows the error rates in each of the frequency classes examined, and the straight line shows an approximate lower bound on these error rates. We can calculate the number of misspelled tokens in each of the frequency classes by

letting T_f = no. of types with frequency f

f = no. of text instances (tokens) of a type with frequency f

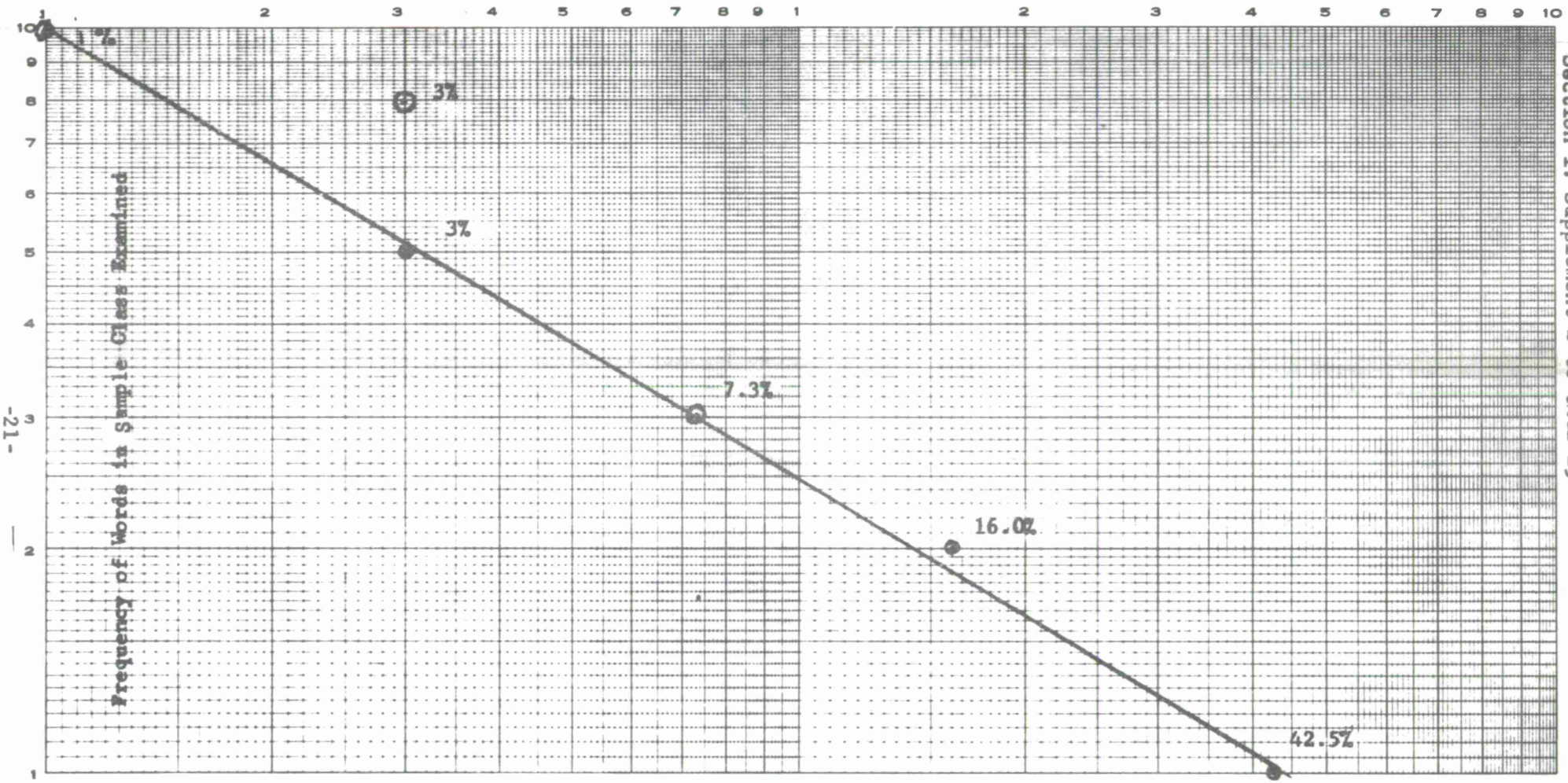
$\mathcal{E}_f$ = error rate of types of frequency f

If we multiply T_f by f we get the number of tokens for words of frequency f. The product

$f(T_f)\ (\mathcal{E}_f)$ gives the desired number,

the number of misspelled tokens contributed by the class of types with frequency f.

By summing from f=1 to 10 we can count an estimated lower bound on the total number of misspelled tokens (for we can assume that the number of errors repeated with frequency greater than 10 is negligible). Since there were ~446,000 tokens in the text, we can estimate the error rate



Percent of Words Considered to be Probable Errors in the Set of Words Having the Designated Frequency

Section I: Supplement 2 to CACL-13

(error rate r in misspellings per 100 tokens) =

$$\frac{\sum_{f=1}^{10} f(\mathcal{E}_f)\ (T_f)}{446{,}000}$$

f	$\mathcal{E}_f$	T_f	$f(\mathcal{E}_f)\ (T_f)$ (approx)
1	.42	12,485	5250
2	.16	2,929	960
3	.07	1,370	288
4	.04	824	132
5	.03	631	100
6	.02	422	50
7	.02	346	50
8	.02	310	50
9	.02	268	50
10	.01	221	22
			6942

An estimate of the error rate is thus

$$\frac{7{,}000}{446{,}000} = 1.5 \text{ errors per 100 words of running text.}$$

As a rule of thumb, we probably have 1/2 misspelled word per abstract on the average.

DISTRIBUTION OF TERM-SET SIZE

FOR THE G.E.-2 AUTO-INDEXED COLLECTION *

A. Introduction

As described in detail in Technical Note CACL-13 , the 10,287 abstracts in Corpus G. E. 2 were automatically indexed by

1. Forming a keyword vocabulary consisting of 1434 word forms corresponding to the 999 most frequent (non-function-word) types in the text of the collection, about 446,000 running words).

2. Indexing an abstract by the set of types appearing in it. Once a type has been assigned to an abstract, the type can be called a term.

The result of this indexing is the production of a matrix C --a binary (10,287 x 999) matrix, each row of which has nonzero entries corresponding to the terms assigned to one of the abstracts. The total number of terms assigned to an abstract by the automatic indexing process is obtained, of course, by counting the number of nonzero entries in the row, and this number is defined to be the term-set size for the given abstract.

The distribution of the term-set size over the collection at hand is an important parameter both for describing the indexing of the collection and for computing (and interpreting) the term association measures that are generated. It will be recalled that the Linear Associative Model** makes provision for a normalization for "document length" (i.e., term-set size) to account for the belief that the cooccurrence of two terms in a "long" document is less weighty than the cooccurrence of two terms in a "short" one. Strict adherence to the model would require accounting for variations in "document length" in computing the associations.

---

* Issued by Paul E. Jones to a limited distribution on May 4, 1965 as Technical Note CACL-16

** See Volume II.

Our computer programs for computing associations do not, strictly speaking, embody this adjustment for variations in term-set size. We weight all cooccurrences equally, and we normalize, in practice, by the cooccurrence total. Use of the computer programs thus embodies the assumption that all term-sets are roughly the same size. In previous work this has been assured by the procedures we have followed (e.g., in G.E.-1 by throwing out all documents with fewer than 7 terms). But in the G.E.-2 collection no such constraints were used; an unfavorable distribution of term-set size could thus conceivably have emerged.

In manually-indexed collections we have reason to believe that there are historical trends in term-set size. B. Dennis of G.E. has reported (private communication to V. E. Giuliano) that early in the formation of the G.E. collection, indexers assigned relatively fewer terms to documents than they did later on. In manual indexing, the average number of terms per document increased with time, presumably because there was increasing need for a finer description of document contents as the collection increased in size.

Although this observation applies to manually-assigned terms, one wonders whether a corresponding trend might not be apparent in the automatically-indexed collection. One can easily speculate that when the collection was first formed, several thousand documents on miscellaneous subjects were on hand and were encompassed in the growing collection before the theme of the collection had a chance to develop. Later, when the aerospace-metallurgy theme became dominant this initial transient could be expected to fade. The administrators of the collection would be expected to be more clearly selective about documents that belonged in the collection. To the extent that a high proportion of the early documents deal with miscellaneous subjects (like agriculture) in which aerospace vocabulary is not employed, they would fail to contain the set of keywords that comes to be repeated frequently later. We would thus expect that the early messages might tend to have fewer keywords in their abstracts than later ones do.

Indeed it is possible to conjecture (pessimistically) that the distribution of term-set size, over the collection as a whole, is bimodal. If early documents are "short" while later ones are "long," this could in principle lead to the presence in the collection of a large number of very short (e.g., 3 terms) documents, together with a very large number of "long" (e.g., 20 terms) documents, and very few in between. Clearly the idea that all term-sets are more-or-less the same size would not be tenable in this case.

B. The Distribution of Term-Set Size as a Function of Accession Number

The availability of the C matrix for Corpus G.E.-2 permitted an analysis to be made of the behavior of term-set size as a function of accession number. For intervals of 200 documents, a count was made of

the number of documents with "length" i (i=1, --- 50)*. No documents of length 0 were encountered, nor were any documents with more than 50 keywords observed.

Figure 1 shows a plot of the distribution of message "length" over the whole collection.** The shape of the distribution is outlined by the line segments connecting the dots. By inspection, the collection as a whole shows a distribution of message length that can be regarded as close enough to normal to dispel worries about bimodality. Nevertheless, there seems to be a slight tendency towards peaking at 20 terms.

Superimposed on this information are plots of the distributions for the first half ot the collection taken alone (bar chart), and for the first 1/4 of the collection (crosses). A slight shift to "shorter" messages among early accessions is seen.

To check the extent of shift in the message length between early and recent documents, Figure 2 shows the distribution for the first 1000 documents overlaid on that for the last 1000 documents. We see that the early segment of documents tends to have one or two fewer terms per document than the late segment does. This difference is not considered sufficiently large to be worth pursuing further.

C. Conclusions

a. For the automatically-indexed collection, the hypothesis that the term-set size exhibits a clear trend to increasing "length" as a function of accession number is partially supported, but the trend is considered negligible. (Note that since we have used a sampling technique, our figures for 1000 G.E.-2 messages represent 7,500 of the original documents. While a major trend is not apparent in our subset, I would not be surprised to see a trend within the first 7,500 G.E. documents.)

b. The distribution of message length is not bimodal over the collection as a whole.

c. The average number of terms per document over the collection as a whole is, accurately:

---

*This data is available on a computer printout. Because it is 50 pages long, it is only summarized here for the collection as a whole.

**The tallies are shown in Table 1.

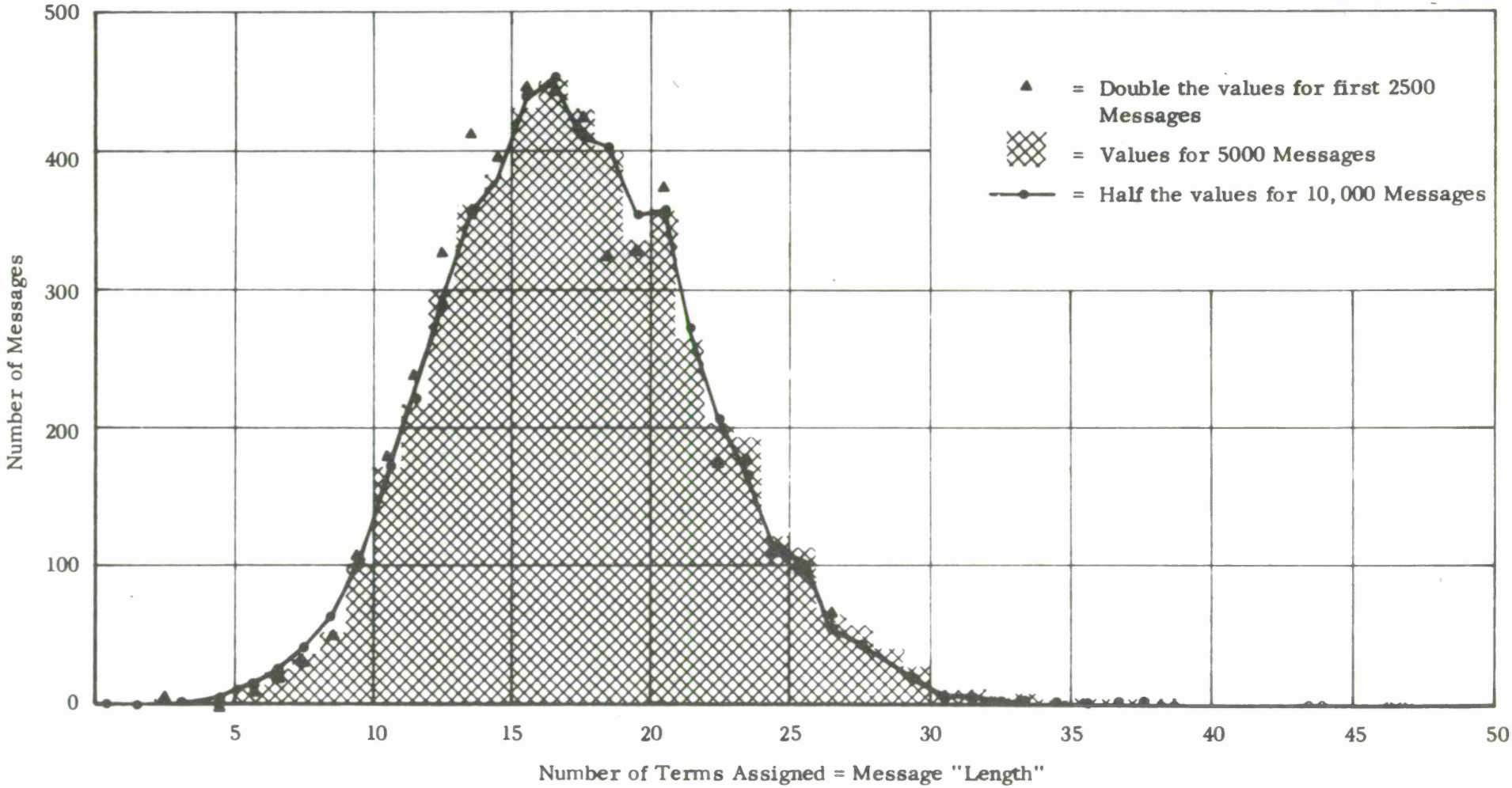


FIGURE I DISTRIBUTION OF AUTO-INDEX TERMS PER MESSAGE FOR THE GE-2 COLLECTION AND ITS INITIAL FRAGMENTS

L Message Length	N No. Docs This Long	LN No. Dt Pairs	L^2N Full No. of TT Pairs
1		0	0
2	2	4	8
3	5	15	45
4	12	48	192
5	35	175	875
6	58	348	2088
7	83	581	4067
8	131	1048	8384
9	207	1863	16767
10	349	3490	34900
11	444	4884	53724
12	574	6888	82656
13	706	9178	119314
14	758	10612	148568
15	874	13110	196650
16	905	14480	231680
17	825	14025	238425
18	807	14526	261468
19	709	13471	255949
20	717	14340	286800
21	541	11361	238581
22	412	9064	199408
23	337	7751	178273
24	230	5520	132480
25	202	5050	126250
26	116	3016	78416
27	89	2403	64881
28	61	1708	47824
29	39	1131	32799
30	17	510	15300
31	13	403	12493
32	7	224	7168
33	8	264	8712
34	3	102	3468
35	2	70	2450
36	2	72	2592
37	3	111	4107
38	1	38	1444
39		0	0
40		0	0
41		0	0
42		0	0
43	2	86	3698
44		0	0
45		0	0
46	1	46	2116
TOTALS - - - - - - -		172016	3105020

TABLE 1

TABULATION OF THE DISTRIBUTION OF MESSAGE LENGTH

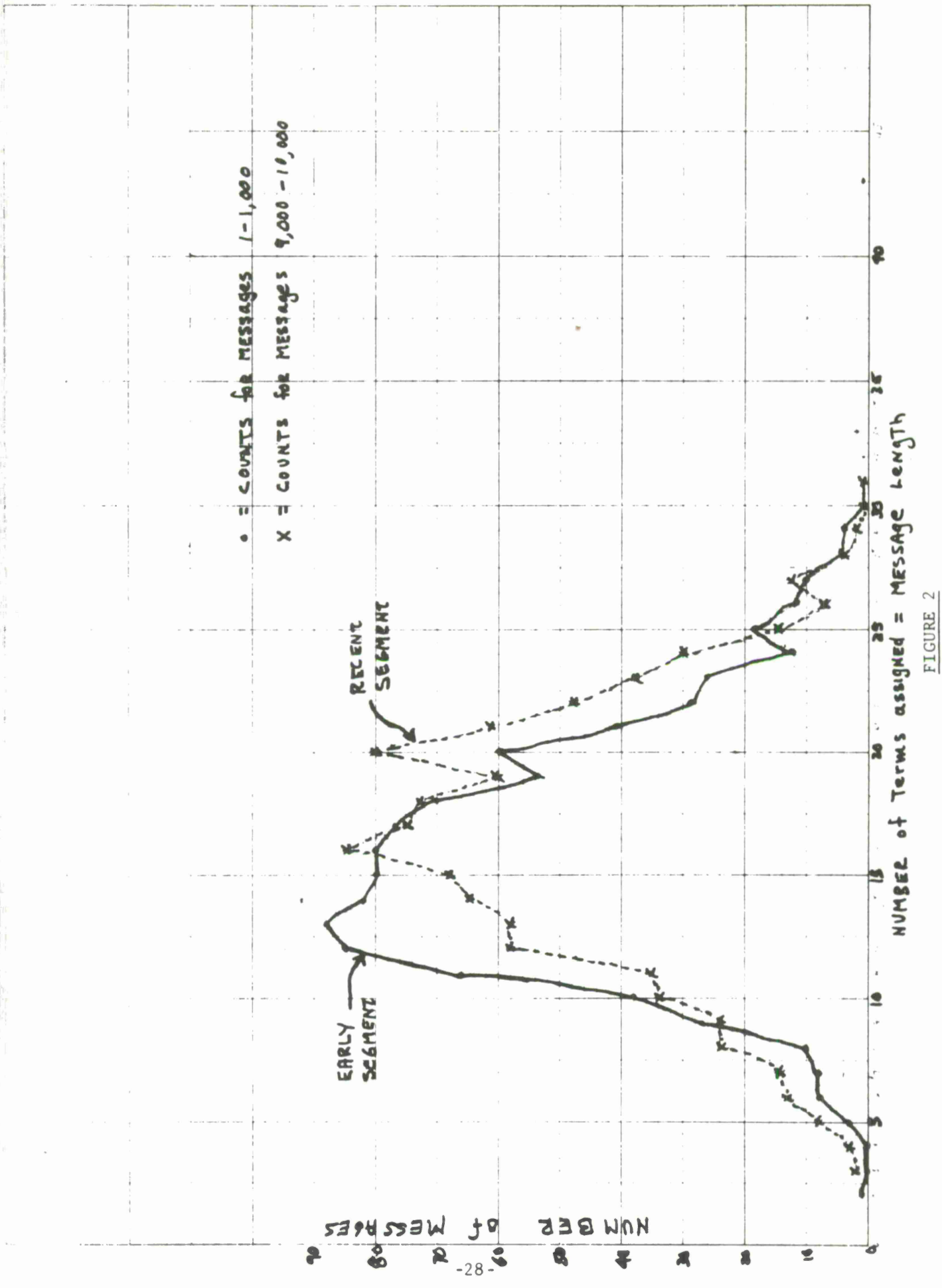

• = COUNTS for MESSAGES 1-1,000
X = COUNTS for MESSAGES 9,000 - 10,000
RECENT SEGMENT
EARLY SEGMENT
NUMBER of Terms assigned = MESSAGE LENGTH
NUMBER of MESSAGES

FIGURE 2

$$\frac{172016}{10287} = 16.72 = \text{aug. terms per document.}$$

d. The total number of term pairs for the collection as a whole is, accurately:

$$L^2N - LN = 2,933,004$$

This corresponds to 285 pairs per document.

SECTION II

COMPARISON OF MANUAL AND MACHINE SELECTED VOCABULARIES

Similarities and Differences of Machine Selected GE 2 Indexing Vocabulary and Manual Uniterm Vocabulary*

The purpose of this Technical Note is to compare certain gross features of the machine-selected G.E.-2 associative indexing vocabulary against those of the UNITERM vocabulary used to index the parent G.E. collection.

I. BACKGROUND

A. Parent G.E. Collection

The G.E.-2 collection consists of 10,289 abstracts indexed by 999 machine-selected "association" terms representing 1,434 singular and plural word forms. The abstracts were selected by sampling from a larger collection which we purchased in 1962 from General Electric Co. This parent collection consists of some 70,000 document surrogates, each consisting of a document number and a set of assigned descriptors. Actual abstracts for only about 45,000 of the documents have been made available to us. These documents were indexed manually by G.E. using UNITERM descriptors; about 4,780 descriptors were used to index the parent collection, with the average number of index terms per document being 12.5. We have available a list of these UNITERMS and their use frequencies in the parent collection; we also have available on magnetic tape all 70,000 surrogates, the 45,000 abstracts and various auxiliary data relating to the parent G.E. collection.

B. G.E.-2 Collection

The 10,289 abstracts G.E.-2 subcollection was selected by machine. The basic procedure used was to select every fourth abstract out of the 45,000 at hand, but skipping those less than six lines long. Since the parent collection was ordered according to accession date, the G.E.-2 subcollection represents a rather uniform sampling in time of the entire parent collection.

---

* Issued on April 30, 1965 to a limited distribution by Janet J. Foster and Vincent E. Giuliano as Technical Note CACL-15. References only have been updated.

Details of preparation of the G.E.-2 corpus have been given in Technical Note CACL-13. The 10,289 abstracts provide roughly 446,000 words of running text. If a non-function word occurs in singular and plural forms a total of at least 56 times, it is used as an "association" index term, and there are 999 such terms. A much larger set of words and word strings may also be used as auxiliary "coordinate" index terms, but these are not of concern here.

II. OVERLAP OF VOCABULARIES

Basically, the matter of concern of this memorandum is the extent to which the two vocabularies in question--the UNITERM vocabulary of the parent collection and the machine-selected G.E.-2 vocabulary--are alike or different. The vocabularies were compared manually to the extent that this is possible without reference to specific documents, and results are summarized here.

A. Inclusion of G. E. -2 Terms in the UNITERM Set

The first question asked was "How many of the G. E.-2 terms are either identical with or close cognates of terms in the UNITERM set?" The 999 G. E.-2 terms were classified into one of the following four categories:

C The G. E.-2 term is either a UNITERM or is the plural of a UNITERM.

X-C The G. E.-2 term is not in the UNITERM list but has the same morphological stem as a UNITERM and is closely related to it in meaning. (Examples: accelerate and acceleration; grow and growth; capable and capability.) Such terms will be referred to here as morphological cognates (listed in Appendix A).

X The G. E.-2 term is neither in the UNITERM list nor a morphological cognate of a UNITERM. (Listed in Appendix B.)

X-G Not otherwise classified. There are 39 unusually short terms in the G. E.-2 list which were not classified because of the difficulty of determining their meaning out of context. These terms contain between 1 and 3 letters, and in many cases are abbreviations of words in the UNITERM vocabulary. These terms could readily be deleted by machine, if desired, on the basis of their length. Percentage figures given in the remainder of this memorandum exclude these 39 terms. (Appendix C)

We accumulated information as to both the number of types and number of token usages of words in the categories, the results being:

TABLE 1

G.E.-2 TERMS INCLUDED IN THE UNITERM VOCABULARY

Category	Type Frequency	% of Types	Token Frequency	% of Tokens
C	730	76%	168,710	84%
X-C	165	17%	22,332	11%
X	65	7%	8,958	5%

Note: Only type X-C, X and X-G tokens were counted in the G.E.-2 collection. Total token usage was estimated by multiplying 10,289 (number abstracts) times 19.5 (estimated number tokens per abstract, based on 15 distinct term types per abstract and recurrence factor of terms being 1.3) giving a total of about 200,000 token usages.

It is of interest that such a high percentage of the G.E.-2 terms are included in the UNITERM vocabulary, particularly considering actual token usages in indexing. The 84% inclusion suggests that a simple machine-derived indexing vocabulary need not be very different in formal makeup than a UNITERM vocabulary consciously selected by indexers. It is of interest to analyze further the machine-selected G.E.-2 terms which are not in the UNITERM vocabulary.

III. ANALYSIS OF DIFFERENCES

A. Morphological Cognates

Most of the UNITERMS are nouns, but many of the morphological cognate X-C terms in the G.E.-2 collection are not. We broke the list of X-C cognate words down according to whether the G.E.-2 term is likely to be primarily a noun, a verb, a verbal (i.e., a participle) or a modifier. The figures for breakdowns into these categories are:

TABLE 2

MORPHOLOGICAL COGNATES

X-C Sub-Category	Type Frequency	% of All Types	Token Frequency	% of All Tokens
Noun	31	3%	3,770	1.9%
Verb	9	1%	1,202	.6%
Verbal	93	10%	12,565	6.3%
Modifier	32	3%	4,795	2.4%
Total X-C	165	17%	22,332	11.2%

The set of 165 type X-C words together with the UNITERM cognates are exhibited in Appendix A; also shown are token frequencies in both the G.E.-2 and the parent collection. In the majority of cases the difference between the text term and UNITERM is primarily one of grammatical form rather than meaning. Detailed inspection of the lists lends very strong credence to the notion that, in the large majority of cases, indexers used a morphological cognate which was an acceptable UNITERM in place of the X-C text term actually used in the text of an abstract. While this notion is both plausible and supported by the data, validation of it will require inspection of the detailed indexing of a sizable number of abstracts; this has not been done so far.

Regardless of how individual abstracts may be indexed, the data of Appendix A does suggest that a concept mentioned in a document can be referred to either by a standard UNITERM (in the case of manual indexing) or by other cognate terms which do appear in text (in the case of machine indexing). Suppose now that one started with a UNITERM vocabulary and attempted to do automatic indexing by means of searching for UNITERMS in text. The data in the above table suggests that at least 11% of the concepts mentioned in the abstracts would not be indexed using this procedure, for 11% of the occurrences of concept-bearing text words would be variants of acceptable UNITERMS, and therefore not recognizable by simple automatic means.

B. Text Words Without Uniterm Cognates

The sixty five G.E.-2 terms which are neither UNITERMS nor which have UNITERM morphological cognates are listed in Appendix B. They are broken down according as to whether they are noun, modifier, verb, verbal or adverb, giving the following data.

Section II: CACL-15

Sub-Category	Type Frequency	% of All Types	Total Token Frequency	% of All Token Occurrences
Noun	19	1.9%	1,922	1.0%
Verb	9	.9%	1,748	.9%
Verbal	13	1.3%	2,786	1.4%
Modifier	21	2.1%	2,305	1.2%
Adverb	3	.3%	197	.1%
TOTAL	65	6.5%	8,958	4.6%

Words of this type (X) are of a general nature and can only be of little value when used singularly for retrieving messages from the present specialized collection. Possibly this is the reason why indexers of the G.E. collection chose not to make them UNITERMS despite their tendency to occur frequently in text. However, such terms could conceivably be of value when combined with other terms having more specific denotations. In any event, the usefulness of such general terms is at present unclear, and will require further evaluation.

APPENDIX A

X-C WORDS

G.E.-2 X-C WORDS	G.E.-2 FREQUENCY	UNITERM COGNATE	PARENT COLLECTION FREQUENCY
I. NOUNS			
acceleration	98	accelerate	(791)
accuracy (ies)	111	accurate	(331)
amplifier (s)	95	amplify	(609)
casting (s)	105	cast	(1023)
coating (s)	425	coat	(1790)
combination (s)	152	combine	(172)
conductivity	143	conduction	(1208)
content (s)	142	contain	(172)
correlation (s)	167	correlate	(220)
curvature	57	curvilinear	(13)
deposition	69	deposit	(572)
derivative (s)	62	derivation	(1223)
diffuser (s)	106	diffuse	(1419)
efficiency (ies)	232	efficient	(1141)
embrittlement	64	embrittle	(318)
environment (s)	115	environ	(561)
feasibility	59	feasible	(68)
forging (s)	99	forge	(784)
formation (s)	159	form	(662)
generation	61	generate	(330)
growth	73	grow	(315)
injection	89	inject	(1080)
instability (ies)	118	instable	(248)
instrumentation	91	instrument	(2017)
loading (s)	205	load	(2567)
presence	94	present	(31)
production	221	produce	(455)
protection	59	protect	(479)
quantity (ies)	73	quantum	(45)

Section II: Appendix A to CACL-15

G.E.-2 X-C WORDS	G.E.-2 FREQUENCY	UNITERM COGNATE	PARENT COLLECTION FREQUENCY
I. NOUNS (Cont.)			
selection	113	select	(184)
separation	113	separate	(595)
II. VERBS			
appear (s)	62	appearance	(7)
consist (s)	71	constitution	(39)
depend (s)	65	dependent	(56)
describe (s)	157	description	(85)
determine (s)	372	determination	(402)
evaluate (s)	76	evaluation	(655)
improve (s)	61	improvement	(98)
include (s)	271	inclusion	(97)
relate (s)	67	relation	(166)
III. VERBALS			
advanced	65	advancement	(175)
aging	85	age	(766)
analyzed	112	analysis	(4333)
applied	274	application	(714)
assumed	133	assumption	(59)
based	335	base	(793)
bending	193	bend	(1259)
boiling	113	boil	(371)
bonded	59	bond	(725)
bonding	71	bond	(725)
calculated	174	calculation	(1966)
calculating	89	calculation	(1966)
carried	72	carrier	(110)
caused	67	cause	(20)
closed	87	close	(343)
combined	103	combine	(172)
compared	280	comparison	(333)
computed	66	computation	(867)

G.E.-2 X-C WORDS	G.E.-2 FREQUENCY	UNITERM COGNATE	PARENT COLLECTION FREQUENCY
III. VERBALS (Cont.)			
computing	65	computation	(867)
conducted	181	conduction	(1208)
conducting	83	conduction	(1208)
consisting	57	constitution	(39)
controlled	100	control	(5202)
cooled	137	cool	(2318)
cooling	318	cool	(2318)
covering	71	cover	(41)
cracking	59	crack	(1126)
curved	63	curve	(1104)
cutting	94	cut	(348)
damping	147	damp	(816)
derived	253	derivation	(1223)
described	396	description	(85)
designed	180	design	(6403)
detailed	73	detail	(22)
determined	347	determination	(402)
determining	187	determination	(402)
developed	404	develop	(1119)
elevated	244	elevation	(170)
employed	75	employment	(29)
established	71	establishment	(14)
evaluated	118	evaluation	(655)
existing	77	existence	(13)
extending	73	extension	(127)
fixed	77	fix	(222)
flowing	59	flow	(8725)
following	83	follower	(17)

G.E.-2 X-C WORDS	G.E.-2 FREQUENCY	UNITERM COGNATE	PARENT COLLECTION FREQUENCY
III. VERBALS (Cont.)			
forced	98	force	(1175)
formed	71	form	(662)
forming	116	form	(662)
generated	64	generate	(330)
heated	119	heat	(7899)
heating	224	heat	(7899)
improved	138	improvement	(98)
included	183	inclusion	(97)
including	162	inclusion	(97)
increased	102	increase	(106)
increasing	77	increase	(106)
indicated	106	induction	(555)
limited	96	limit	(558)
loaded	67	load	(2567)
manufacturing	'72	manufacture	(660)
measured	199	measure	(581)
melting	131	melt	(816)
mixing	99	mix	(785)
moving	59	movement	(148)
observed	101	observe	(10)
operating	229	operate	(489)
past	89	pass	(36)
performed	88	perform	(1285)
predicted	69	prediction	(305)
prepared	66	preparation	(275)
presented	62	present	(31)
processing	124	process	(1037)
produced	191	produce	(455)

G.E.-2 X-C WORDS	G.E.-2 FREQUENCY	UNITERM COGNATE	PARENT COLLECTION FREQUENCY
III. VERBALS (Cont.)			
producing	71	produce	(455)
proposed	96	proposal	(125)
reinforced	97	reinforce	(407)
related	148	relation	(166)
relating	62	relation	(166)
reported	107	report	(180)
resulting	68	result	(83)
reviewed	82	review	(305)
selected	111	select	(184)
simulated	57	simulating simulation	(38) (430)
solved	64	solving	(14)
starting	67	start	(426)
studied	231	study	(561)
supported	94	support	(398)
taken	59	take	(29)
used	740	use	(79)
using	572	use	(79)
varying	85	vary	(429)
welded	82	weld	(2017)
IV. MODIFIERS			
analytical	181	analysis	(4333)
annular	80	annulus	(459)
applicable	123	application	(714)
available	134	availability	(61)
axisymmetric	66	axisymmetry	(290)
basic	243	base	(793)
capable	71	capability	(75)

Section II: Appendix A to CACL-15

G.E.-2 X-C WORDS	G.E.-2 FREQUENCY	UNITERM COGNATE	PARENT COLLECTION FREQUENCY
IV. MODIFIERS (Cont.)			
continuous	102	continuation	(224)
cylindrical	198	cylinder	(1406)
digital	91	digit	(643)
dimensional	431	dimension	(1057)
experimental	881	experiment	(827)
flexural	58	flexure	(266)
gaseous	78	gas	(5347)
German	72	Germany	(55)
magneto hydro	74	magneto	(541)
mathematical	86	mathematics	(816)
metallic	91	metalloid	(24)
metallurgical	71	metallurgy	(1074)
operational	102	operation	(887)
optical	74	optic	(398)
partial	97	part	(402)
protective	59	protect	(479)
random	60	randomness	(140)
rectangular	114	rectilinear	(21)
relative	120	relativity	(40)
significant	82	significance	(10)
spherical	88	sphere	(494)
structural	309	structure	(2819)
theoretical	395	theory	(1662)
typical	88	type	(127)
useful	76	use	(79)

APPENDIX B

X WORDS

(N = noun; M = modifier; V = verb; VB = verbal and A = adverb)

G.E.-2 X WORDS		G.E.-2 FREQUENCY	G.E.-2 X WORDS		G.E.-2 FREQUENCY
advantage(s)	N	142	good	M	126
agreement	N	88	graphical	M	59
amount(s)	N	91	greater	M	75
arrangement(s)	N	57	importance	N	62
associated	VB	133	important	M	77
attempt(s)	N	70	involved	VB	80
attention	N	62	involving	VB	69
best	M	94	known	M	100
better	M	70	made	V	518
cent	N	76	make(s)	V	70
certain	M	165	necessary	M	86
columbium	N	99	need(s)	N	80
complete	M	114	nonlinear	M	141
consideration	N	183	note(s)	N	68
considered	VB	276	now	A	80
discussed	VB	497	obtained	VB	545
discuss(es)	V	109	obtaining	VB	64
discussion(s)	N	294	obtain(s)	V	112
due	M	174	occur(s)	V	100
encountered	VB	65	output(s)	N	99
expression(s)	N	115	particularly	A	58
feature(s)	N	106	particular	M	107
found	VB	263	permit(s)	N	89
further	M	76	possibility(ies)	N	72
give(s)	V	166	possible	M	233
given	VB	494	previously	A	59

Section II: Appendix B to CACL-15

G.E.-2 X WORDS		G.E.-2 FREQUENCY	G.E.-2 X WORDS		G.E.-2 FREQUENCY
previous	M	71	showing	VB	65
principal	M	76	shown	V	246
provide(s)	V	203	show(s)	V	224
recent	M	103	subjected	VB	162
same	M	119	subject(s)	N	69
satisfactory	M	83	suitable	M	156
showed	VB	73			

APPENDIX C

XG WORDS

G.E.-2 XG WORDS	G.E.-2 FREQUENCY	G.E.-2 XG WORDS	G.E.-2 FREQUENCY
AD	205	K	92
AF33	94	LA	69
AF	249	L	68
AL	115	M	7364
B	102	N.	102
CO	65	O	84
CR	125	PCT	241
C	382	P	95
DEG	396	PSI	98
DER	85	QPR	62
DE	163	RE	90
DES	103	R	76
D	201	SEC	69
E	59	S	86
F	402	T	85
FT	81	VOL	62
H	57	V	164
III	88	W	124
II	200	X	241
J	71	Total	12515

SECTION III

VOCABULARY DISTRIBUTION STUDIES

Token Frequencies and Entropy Calculations for the GE 2 Collection*

A. Introduction

Detailed knowledge of the statistical parameters of the G.E.-2 collection is of considerable importance to the task of extrapolating results obtained with this collection to others that might be "similar". In other Technical Notes, the report to the USAF (ESD - TR-66-405) and elsewhere, a number of such observations and measurements upon the collection have already been recorded as a byproduct of one investigation or another. Continuing this process of determining the values of the collection parameters believed to be important, this note is primarily concerned with the calculation of the uncertainty statistics for the word strings that were used in the collection. We capitalize on the somewhat unusual opportunity offered to us to determine the values of H(x),..., H(w, x, y, z) for a body of scientific text of about 500,000 words as an incidental byproduct of other work.

Because our procedures for gathering the string statistics affect the way in which the uncertainty measures may be interpreted, we first review these procedures briefly in Section B. The entropy calculations are then recorded for strings of length 1, 2, 3, and 4 in Section C. The discussion in Section D is devoted to developing estimates of conditional entropy for the collection. It provides a partial description of the corpus using information-theoretic terms. Finally, we turn attention to the subset of the corpus that was "underlined" by automatic indexing. The entropy attributable to the words used in automatic indexing (the G.E.-2A vocabulary) is determined and it is shown that about 51% of the total entropy is contributed by these words.

---

* By Paul E. Jones. Not previously issued.

Section III: Technical Note

B. Gathering String Statistics*

1. The Collection

The data base on which the present experimental observations were made is the GE-2 collection, consisting of 10,287 abstracts selected from a GE parent collection of approximately 45,000 abstracts. We excluded abstracts that were "short" (e. g., those which were very brief) but the procedures for selection were otherwise random, involving picking every third one. Each abstract consists, for the present purposes, of a title and the text of the abstract that was given.

2. Word Form Tokens

In processing this collection by computer, the text of each abstract was considered to be a string of word form tokens; i.e., basically strings of characters separated by the space symbol. There were some minor variants to the use of space as a word form token separator. Specifically, hyphenation in the creation of composite words was ignored: for example, META-THEORY was regarded as two word forms META and THEORY in successive positions. On the other hand, end-of-line hyphenation was recognized, and broken words were glued together. Naturally, there were instances where end-of-line hyphenation corresponded to composite word creation, and in these (rare) instances, the words were (improperly) glued together. There were also many instances in which, due to transcription errors, this end-of-line hyphen was not present. In such cases, extra words came to be generated (e.g., DOCT-OR could yield the word form DOCT if the hyphen were omitted.)

By and large, however, the definition of word form token corresponds accurately to the natural segmentation of text using spaces. But note for completeness that numbers (like 0.02) count as word forms and that symbols (like W for Tungsten) do, too. Moreover, note that the Roman numeral I is not distinguishable from the pronoun of the same form, etc.

---

* The programs used for this purpose were prepared and run by Miss Joyce Mehring. For a detailed description, see Technical Note CACL-13.

Section III: Technical Note

The comma (,) and the period (.) punctuation marks are ignored. Each abstract was therefore "seen" as a string of m tokens, none of which was a period or comma. All our statistics are based on this view of the text.

3. String Generation and Counting

Based on this view, string statistics were obtained by

a. "Passing a four-word window" over the m word string for every abstract and recording each such four-word string separately on an output tape.

Example: If the abstract began with the string of word forms: a b c d e f g ...

we would generate

abcd

bcde

cdef

defg

etc.

One string is generated for each text position at the beginning of the abstract. In order to permit this correspondence to continue at the end of the abstract, we added up to three "dummy" positions (blanks) at the end. Thus, if the end of the abstract string were

... u v w x y z , we would generate:

uvwx

vwxy

wxyz

xyz^

yz^^

z^^^

b. The set of such four-word strings was alphabetically sorted to group recurrences of the initial substrings of the four-word strings. This is the "Master Context List".

c. These initial substrings were counted to yield string statistics as described below.

Section III: Technical Note

C. Frequency Data and Entropy Calculations for Strings

1. Single Word Forms

a. Frequency Data for Single Words

The Master Context List contains a four-word string for each position of the text; the word form that appeared in a given position in the text is the first constituent of the corresponding four-word string. To count frequencies of single word forms, these "first constituents" in the sorted Master Context List were tallied. As a result, word "types" were constructed and the frequency of usage of each was obtained.

These types (with frequency) were then arranged in frequency order and a second summary was made by counting the number of types occurring with each frequency. This process yielded a report stating, for each distinct frequency that was observed, the number of types that occurred with that frequency. For example, there was

1 type with frequency 30170
1 type with frequency 16622
⋮

1370 types with frequency 3
2929 types with frequency 2
12485 types with frequency 1

Using these figures, it was possible to count the total number of tokesn in the text, by accumulating the total number of tokens accounted for by each line of the above summary. Thus, the first line accounts for 30170 tokens, the next to last line accounts for 2 x 2929 tokens, and the last accounts for 12,485 tokens. The grand total was 446,097 tokens, and this is the total number of text positions.

There were 23,505 different types (i.e., distinct word forms) recognized by the formal procedures described in Section A. Thus, this set of types includes numerals,

misspellings, parts of improperly broken words; i.e., exactly as they appeared in the text. This set of 23,505 types is taken as the symbol vocabulary of the collection.

b. Entropy Calculations for Single Words

The entropy H(x) of this symbol vocabulary was computed based on the following procedure:

(1) From each line of the summary, we know a distinct frequency, f_i, for some set of types. A type of this frequency contributes

$$h_i = -\left(\frac{f_i}{446,097}\right) \log_2 \left(\frac{f_i}{446,097}\right)$$

to the total entropy H(x).

(2) If there are n_i types that occurred with that frequency f_i, in aggregate they contribute $n_i h_i$ to the total entropy H(x).

(3) The total entropy H(x) was obtained by cumulating the $n_i h_i$ for the lines of the summary report. The calculated value was:

$$H(x) = 10.36 \pm 0.01 \text{ (Estimated error) bits per word.}$$

The estimated error is probably larger than it should be. A more detailed analysis would take note of the fact that 10 digits were used in the mantissa of floating-point calculation in 1401 FORTRAN 2 and that the major source of error is probably the LOG routine provided in the system library--whose error properties are not known to us at this time. Any systematic error in that routine--e.g., a value slightly too low--would be accumulated 23,505 times by this process.

2. Two Word Strings

a. Frequency Data for Two-Word Strings

Almost exactly the same procedure was followed with two-word strings as that used for single words. However, because there were so many one-occurrence two-word strings, these

were not actually counted. Rather, all the two-word strings that occurred more than once were counted. As before, types were produced, arranged by frequency and summarized. The summary text was of the same form; e.g.,

1 type with frequency 2331

1 type with frequency 1469

⋮ ⋮

8639 types with frequency 3

15877 types with frequency 2

166,911 types with frequency 1.

b. Entropy Calculation for Two-word Strings

The entropy of the pair vocabulary is easily calculated using the same procedure as before. That is, we would readily regard any pair that occurs with frequency f_i to contribute

$$h_i = \left(\frac{f_i}{N}\right) \log_2 \left(\frac{f_i}{N}\right)$$

to the entropy Shannon calls H (x,y). Had our entire collection of abstracts been regarded as a single long string, this would indeed by the correct procedure. But in the procedure we followed for moving a window over the text, we introduced a dummy (blank) at the end of each message and regarded the collection as composed of 10,287 shorter units. It follows that we have generated pair types of the form "z ∧", that in aggregate occur 10,287 times. Some of them probably occur quite frequently (e.g., 30 times) reflecting some propensity to end abstracts with the word z . We do not know which types are of this "terminal" kind; accordingly, we must be careful in interpreting the value of H(x,y) obtained from the formula.

Section III: Technical Note

We regard our symbol source as something which emits a special terminating symbol (dummy) at the end of every message. To identify this notion of source (in contrast to the case of single words where there was no need to consider the dummy symbol) we call the pair entropy $H^{1D}(x,y)$ to note the fact that one Dummy is emitted per message.

The calculation of $H^{1D}(x,y)$ is performed by computing the probability $P^{1D}(xy)$ of each pair type using

$$P^{1D}(xy) = \frac{f^{1D}(xy)}{446,097}$$

where $f^{1D}(xy)$ is the pair frequency given in the summary list and 446,097 is the total number of such pairs in our sample of symbol strings emitted by the 1D sources.

The value of $H^{1D}(x,y)$ was calculated for the collection using the same procedure previously described. The result was:

$$H^{1D}(x,y) = 16.26 \pm .01 \text{ bits per digram}$$

3. Three-Word Strings

a. Frequency Data for Three-word Strings

The same procedure was used for tallying three-word frequencies as that described for pairs, except that strings with frequency 2 and 1 were not counted. Nor was it possible to deduce the exact members of strings with these frequencies as was possible in the case of two-word strings.

It was possible, however, to calculate bounds on the entropy by considering the extreme way the remaining

(446,097 - 64,412) = 381, 687 residual 3 strings

could be distributed over 2-occurrence and 1-occurrence types. The two extreme cases are shown below.

Case A: Suppose the three-word strings in question are all at frequency 1, i.e.,

0 types with frequency 2

381,687 types with frequency 1

total types = 381,687

Case B: Suppose as many as possible of the three-word strings occur with frequency 2 (and as few as possible with frequency 1). Because there are 166,911 two word types AB that occur only once, we know there are 166,911 triplet types ABC that also occur only once. Thus Case B results in

107,388 types with frequency 2

166,911 types with frequency 1

b. Entropy Estimate for Three-word Strings

Using the same procedure previously described under the discussion for two-word strings, we obtained the following values of H^{2D} (x, y,z) for the assumed extreme distributions of low frequency types.

H^{2D} (x,y,z) = 18.34 (all residual strings taken as f=1)

H^{2D} (x,y,z) = 17.86 (as many residual strings taken to have f=2 as possible)

Thus, H^{2D} (x,y,z) = 18.10 $\pm$ 0.24

4. Four-Word Strings

a. Frequency Data for Four-word Strings

The situation with four-word strings is exactly like that for three-word strings. The number of types that occurred with frequency $\geq$ 2 was not recorded. The residual (446,097 - 14,855) = 431,242 four-word strings could be distributed in the same extreme ways as before:

Four-word strings

Case A: Assume all the types have frequency 1

431,242 types with frequency 1

0 types with frequency 2

Case B: Assume as many as possible of the residual types have frequency 2. Arguing as before, we know there are at least 166, 911 triplets with frequency 1 and hence at least that many quadruplets. Thus the extreme case is

166,911 types with frequency 1

132,165 types with frequency 2

b. Entropy Estimate for Four-word Strings

The entropy calculations were carried out for both cases as usual. The contribution due to all the quadruplets with $f \geqslant 3$ is, we note, only 0.54.

Case A: $H^{3D}(w,x,y,z) = 18.7$

Case B: $H^{3D}(w,x,y,z) = 18.1$

Thus, $H^{3D}(w,x,y,z) = 18.4 \pm .4$ bits per quadruplet

D. Conditional Entropy Estimates

1. General

The ordinary procedure for determining conditional entropy cannot be applied at once to the data we have gathered. As pointed out in Section C-2b, we are not treating the text as one long fragment but as 10,287 short pieces. We have introduced "dummy" symbols between the messages, and the presence of these dummies is a complication that must be dealth with. To exhibit the difficulty most clearly, we need only notice that $H^{1D}(x,y)$ was obtained from a source among whose symbols a dummy was present, whereas $H(x)$ was obtained from a source that produced no dummies. Thus, even though the entropy values calculated in Section C were all based on the same corpus, we -- in a limited sense --treated it as four distinct sources in the information theoretic sense.

The formulas for conditional entropy, e.g.,

$$H(x,y) - H(x) = H_x(y)$$

cannot be used without giving some attention to this point.

2. Approximations and Estimates

Suppose we treat the source as a 1D# source, i.e., the corpus consists of the texts of the messages with a single dummy space between each. We need to adjust $H^{1D}(x,y)$ and $H(x)$ to correspond to this source in order to find $H^{1D\#}\ x\ (y)$. The value of $H^{1D\#}(x)$ however, is very closely related to $H(x)$; one more symbol (the

dummy) has been added with frequency 10,287 and the total length of the sample has been increased by 10.287. The addition of the n new symbol raises H by

$$- \frac{10,287}{446,097} \log_2 \frac{10,287}{446,097} = .0229 \log_2 = 0229 = .12$$

The increased length of the sample changes all the P_i from expressions of the form $\frac{f_i}{446,097}$

to expressions of the form

$$\frac{f_i}{446,097 + 10,287}$$

Accordingly, all the $P_i \log P_i$ aggregated to form H (x) are slightly reduced in forming $H^{1D\#}(x)$. Since the two effects tend to balance out to some extent we can estimate

$$H^{1D\#}(x) = (H(x) + .05) \pm .05$$

The same kind of argument needs also to be applied to adjust the computed value of $H^{1D}(x,y)$. In tallying the strings, we did not count 10,287 pairs of the form ⟨dummy, first word of message⟩. Also, we are now considering that the text includes the dummies, so that the P_i are all reduced.

Again, the effects are counteracting, and we deal with it by increasing the error bounds on $H^{1D}(x,y)$:

$$H^{1D\#}(x,y) = H^{1D}(x,y) \pm .05$$

We can now compute the conditional entropy for the 1D# source.

$$H_x^{1D\#}(y) = H^{1D\#}(x,y) - H^{1D\#}(x)$$
$$= 16.26 \pm .06 - (10.36 \pm .01 + .05 \pm .05)$$
$$H_x^{1D\#}(y) = 5.95 \pm .12$$

The foregoing discussion outlines the factors that must be considered before the uncertainty figures reported here can be compared and contrasted with those for other collections used as sensitive parameters in theoretical work. It is clear, however, that the

various ways of interpreting the nature of the symbol source cause only minor fluctuations in the entropy values. For many informal purposes, these variations can be ignored.

With this view in mind, we estimate

$$H_y(z) = 5.9 \pm .2 \text{ bits}$$

$$H_{xy}(z) = 18.1 \pm .3 - (16.3 \pm .1)$$

$$= 1.8 \pm .4 \text{ bits}$$

The cumulative errors are too great to permit us to estimate $H_{wxy}(z)$.

3. Contribution to H(x) of GE-2A Automatic Indexing Vocabulary

The subset of 1434 word forms making up the GE-2A vocabulary was identified and their contribution to H(x) was calculated using the formula

$$- \frac{f_i}{446,097} \log_2 \frac{f_i}{446,097}$$

The value obtained was $5.25 \pm .01$ bits for a total of 219,913 tokens or 49.30% of the text. Since $H(x) = 10.36 \pm .01$ bits, 51% of the total is contributed by our GE-2A vocabulary.

Section III: Supplement to CACL-30

ZIPF'S LAW AND HERDAN'S LAW OF SOLIDARITY *

A. Introduction

Zipf's Law for vocabulary distribution states that the rank of a word type times its frequency of usage in a sample of text is (approximately) a constant. Herdan** arranges the same data differently and claims that the number of word types occurring with frequency n is predicted by the Waring distribution. (See TN-CACL-30) He points, with especial interest, at the gradient of the curve, expressed as the ratio of the number h_{n+1} of types occurring with frequency n+1 to those occurring with frequency n. This ratio is given by the Waring distribution to be

$$\frac{h_{n+1}}{h_n} = \frac{a+n-2}{x+n-1} \qquad (1)$$

where a and x are constant parameters of the distribution. Herdan considers this feature of the Waring distribution - the regularity of the ratio of successive terms - to be quite significant. (p.89)

> This relation between successive terms of the series remains unaltered (invariant) despite the change in numerical values with sample size. The gradient is thus an invariant, epitomizing the system of solidarity among the vocabulary items.

On page 91 he writes

> This means that the accumulation of vocabulary items in the various classes of the variable follows some sort of solidarity mechanism. Such a mechanism is implied by the Waring distribution, representing a gradient of probabilities, each of which stands in a definite relation to the preceding one, which constitutes what we have called the invariance of the vocabulary distribution function.

And later

> "Considering that the latter (gradient of the frequencies in the successive class intervals of the variable described by the Waring distribution) is brought about by the solidarity of the system of vocabulary, I propose to call it a Law of Solidarity..."

---

* Issued on May 10, 1966 to a limited distribution as a Supplement to Technical Note CACL-30 by Paul E. Jones, Jr.

** Herdan, G., Quantitative Linguistics, Butterworths, London, 1964.

Section III: Supplement to CACL-30

This note is devoted to demonstrating that a corpus which satisfies Zipf's Law has a "Law of Solidarity" of almost identical form. Moreover, when this result is coupled with the empirical result presented below (which showed that a Zipf plot of the Herdan-Waring distribution yields a straight line for the data we studied), the conjecture that the Herdan-Waring distribution is very similar to Zipf's is reinforced.

B. The Corresponding Gradient in a Zipf Distribution

Consider a sample of text containing a total of T distinct types, of which h_1 occur exactly once. Let h_n in general denote the number of types that occur with frequency n. Let r_n designate the "rank" of those types that occur with frequency n. (Since there may be many types occurring with frequency n, we adopt the convention that by "rank" r_n we mean the largest rank assigned to the set of words with frequency n. On a Zipf plot, then, we consider the right hand edge of the descending steps to be the point which defines "rank".)

Consider a sample in which Zipf's Law holds -- using the foregoing definition of "rank" for types with the same frequency.

Because we know there are T types in all, we know that $r_1=T$, that $r_2=T-h_1$, and in general that

$$r_n = T - \sum_{i=1}^{n-1} h_i$$

$$r_{n+1} = r_n - h_n$$

But if Zipf's Law holds, then

$$n.r_n = \text{const.} = (n+1)\, r_{n+1} = (n+1)\, r_n - (n+1)\, h_n$$

$$= n.r_n + r_n - (n+1)\, h_n$$

$$\therefore \quad h_n = \frac{1}{(n+1)}\, r_n$$

Similarly,

$$h_{n+1} = \frac{1}{n+2}\,(r_n - h_n) = \frac{1}{n+2}\left(1 - \frac{1}{(n+1)}\right) r_n = \frac{n}{n+2}\, h_n$$

Thus,

$$\boxed{\frac{h_{n+1}}{h_n} = \frac{n}{n+2}}$$

which is a Law of Solidarity for the Zipf distribution comparable to Herdan's version. (1)

Obviously, setting a=2 and x=3 makes the two gradients identical. It would be well to look next at whether a and x can depart significantly from these values in Herdan's development.

c. Conclusion

Herdan may be reading too much into the significance of the Waring distribution's regular gradient. I find it difficult to see what he is so excited about in view of the above result.

Work continues in an attempt to place the Waring distribution in a form that makes it apparent what the corresponding rank-frequency plot looks like.

FITTING THE HERDAN-WARING DISTRIBUTION TO THE VOCABULARY USAGE DISTRIBUTION IN THE GE-2 CORPUS*

A. Introduction

Herdan** is critical of the Zipf model for vocabulary distribution. He prefers rather to work with the number t_n of types that occur with frequency n in the text. For the GE-2 collection we have such data in a table

Number of Types	Frequency
1	30170
1	16222
⋮	⋮
2929	2
12485	1

Herdan suggests that the Waring distribution is an appropriate function to fit to this vocabulary distribution. This note represents a test of Herdan's claim as applied to the GE-2 message collection.

Waring's expansion for 1/x-a is reported by Herdan to go back to the 18th century. This expansion is written (we have corrected the misprints):

$$\frac{1}{x-a} = \frac{1}{x} + \frac{a}{x(x+1)} + \frac{a(a+1)}{x(x+1)(x+2)} + \cdots$$

---

* Issued on May 9, 1966 to a limited distribution as Technical Note CACL-30 by Peter R. Bono and Paul E. Jones, Jr.

** Herdan, G. Quantitative Linguistics. Butterworth, London, 1964. (See especially p. 85 ff.)

and is convergent for $x > a > 0$. If we multiply by (x-a), the series on the right sums to unity, and Herdan asserts that the n^{th} term may be interpreted as P_n, the fraction of the total vocabulary T (i.e., the proportion of all the types in the sample) that are found occurring in the text with frequency n.

Given the correct choices of x and a, Herdan claims that the Waring distribution should fit observed data, like that in the table above that shows the number of types with frequency of occurrence = n. The key to fitting the Waring distribution to the sample data is to choose values of x and a. For this purpose Herdan exhibits the following procedure.

B. Procedure

1. Notation

Let $\bar{x}$ be the average frequency of occurrence of a word; i.e.,

$$\bar{x} = \frac{N}{T}$$

where N = number of running words (446097 in GE-2)

T = number of different types (23505 in GE-2)

and p_1 = fraction of vocabulary of types which is accounted for by words which occur only once (i.e., by 12485 types in GE-2).

p_n = fraction of types occurring with frequency n.

2. Procedure

Herdan provides tables which can be entered using $\bar{x}$ and p_1 above. These tables provide the value p_n (n=2, 3,...,50). So they yield the expected number of types with frequency n if we multiply $p_n \cdot T$.

3. Experiment at fitting

We first attempted to follow Herdan's procedure using our data

and to determine the fit between predicted and observed values. For the GE-2 collection we calculate

$$\bar{x} = \frac{446097}{23505} = 18.9788 \qquad \text{mean frequency}$$

$$p_1 = \frac{12485}{23505} = .53116 \qquad \text{fraction of words with f=1}$$

and the desired predicted values of p_i were obtained from interpolation in Herdan's table on Pages 267-8.

The product $P_i \cdot T$ together with the observed number of types with frequency i, are plotted against i(x-axis) in Figure 1. The figure is cut apart for ease of printing.

The fit was considerably better than had been expected. The two curves have very much the same shape. Because of the fitting procedure, the two curves come together (off-scale) at frequency = 1. On the other hand, the differences between predicted and observed numbers of types is substantial for low frequencies > 1. This causes such large contributions to x^2 that we did not consider it worthwhile to complete the details of the test.

In view of the fitting procedure's dependence on the number of types that occur with frequency 1, we recalled that many of these types counted by the machine were actually misspelled words. Estimates of the number of types in the frequency 1, 2,...,10 ranges which were occurrences or repeated occurrences of misspelled words had previously been obtained (Technical Note CACL-13, Supplement). We wondered whether the blind counting we had employed might not be the reason of the inexact fit in Figure 1.

Accordingly, we -- in effect -- threw out all instances of misspelled words from the running text of the GE-2 corpus and treated the remaining

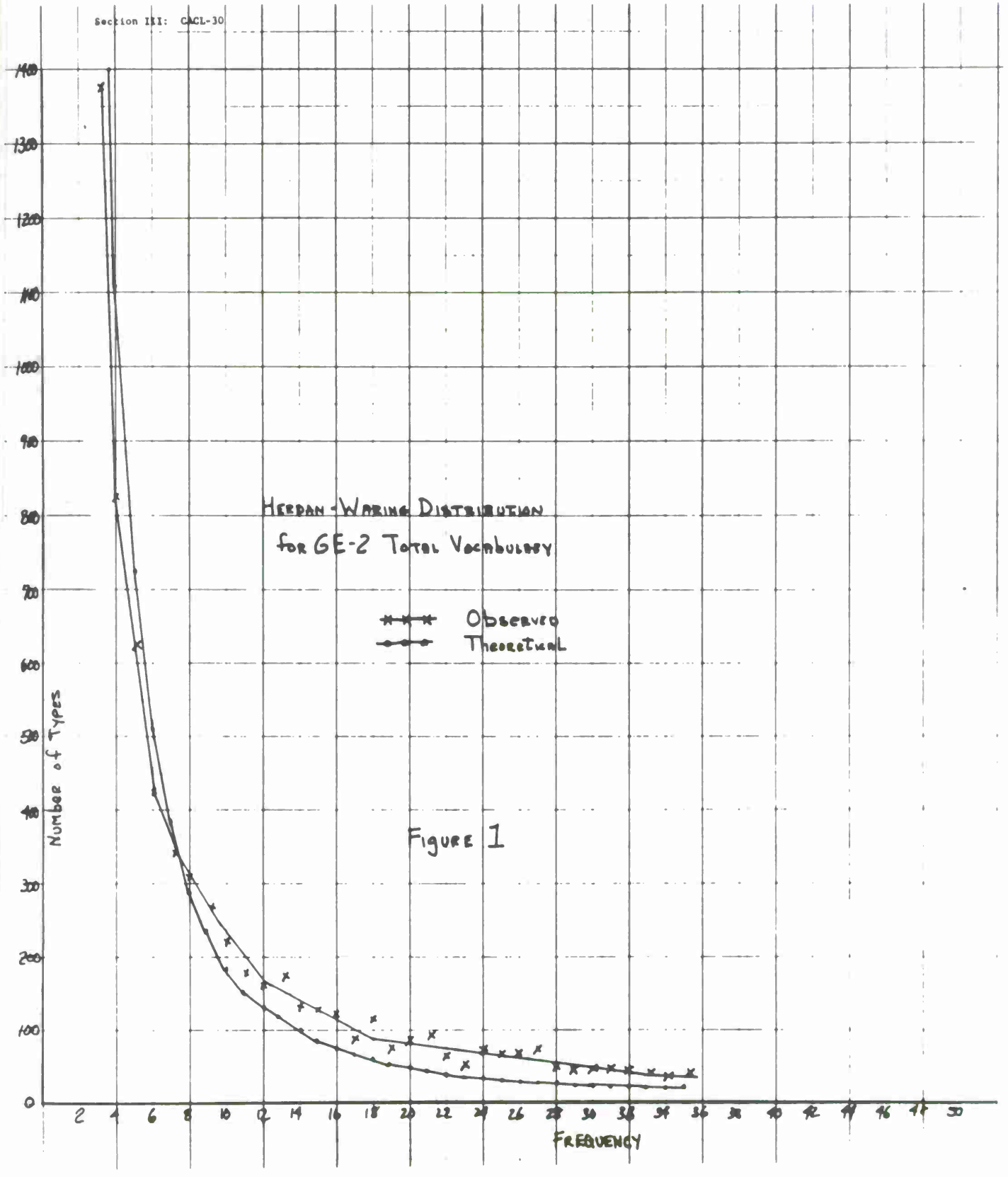

HERDAN-WARING DISTRIBUTION
for GE-2 Total Vocabulary
Observed
Theoretical
Number of Types
FREQUENCY

Figure 1

text as if it had been the original sample. This was accomplished by

(a) subtracting, from the observed number of types with frequency i, the estimated number of misspelled types with that frequency (for i=1,...,10). The resulting new values were

frequency i	Observed Number of Types	Estimated % Misspelled	Types Misspelled	Corrected Number
1	12485	42.5%	5306	7179
2	2929	16.0%	469	2460
3	1370	7.3%	100	1270
4	824	4.5%	37	787
5	631	3.0%	19	612
6	422	2.3%	10	412
7	346	1.8%	6	340
8	310	1.4%	5	305
9	268	1.2%	3	265
10	221	1.0%	2	219
		Total	5957	

(b) The total number of types T was reduced by the number of types misspelled to yield 23505 - 5957 = 17548.

(c) The total number of tokens in the text was reduced by the number of misspelled words "thrown out". (See Technical Note CACL-13, Supplement, last page for determining the estimate). The total = 6976 words misspelled; hence 439121 is the reduced number of tokens in the text.

Because Herdan's tables did not cover the interval into which these numbers fell, a small program was prepared to compute the Waring distribution. We checked the previous set of table values used in develop-

ing Figure 1, and checked Herdan's computation of the distribution on Page 87. Both were correct within the limits of accuracy of the published tables.

Figure 2 shows the plot for the reduced GE-2 collection with misspellings "thrown out". A slight improvement is observed.

C. Conclusions

1. The Herdan-Waring distribution reflects with moderate accuracy the observed distribution of the number of types which have occurrence frequency n .

2. The exclusion of misspelled words from the text in an effort to improve the fit resulted in a modest improvement, but this improvement is less interesting than the accuracy of the fit achieved without any intervention.

3. It would be valuable to determine this distribution, when converted into a predictor of the rank-frequency curve, would account for or give clues about the bow in the Zipf curve. This conjecture will be treated further in subsequent notes. Preliminary evidence suggests that the Herdan-Waring distribution for a sample produces a vocabulary distribution that closely resembles Zipf's, at least at the right hand end of the rank frequency plot. Figure 3 shows the results of plotting the Herdan-Waring values in Zipf form for the two sets of text parameters we ran through the program. Curve A is for the GE-2 collection discussed in this note. Curve B is for the sample of Pushkin to which Herdan applies his procedures (Page 87) and which we used to check the program. A straight line of slope -1 is included as usual to aid the eye.

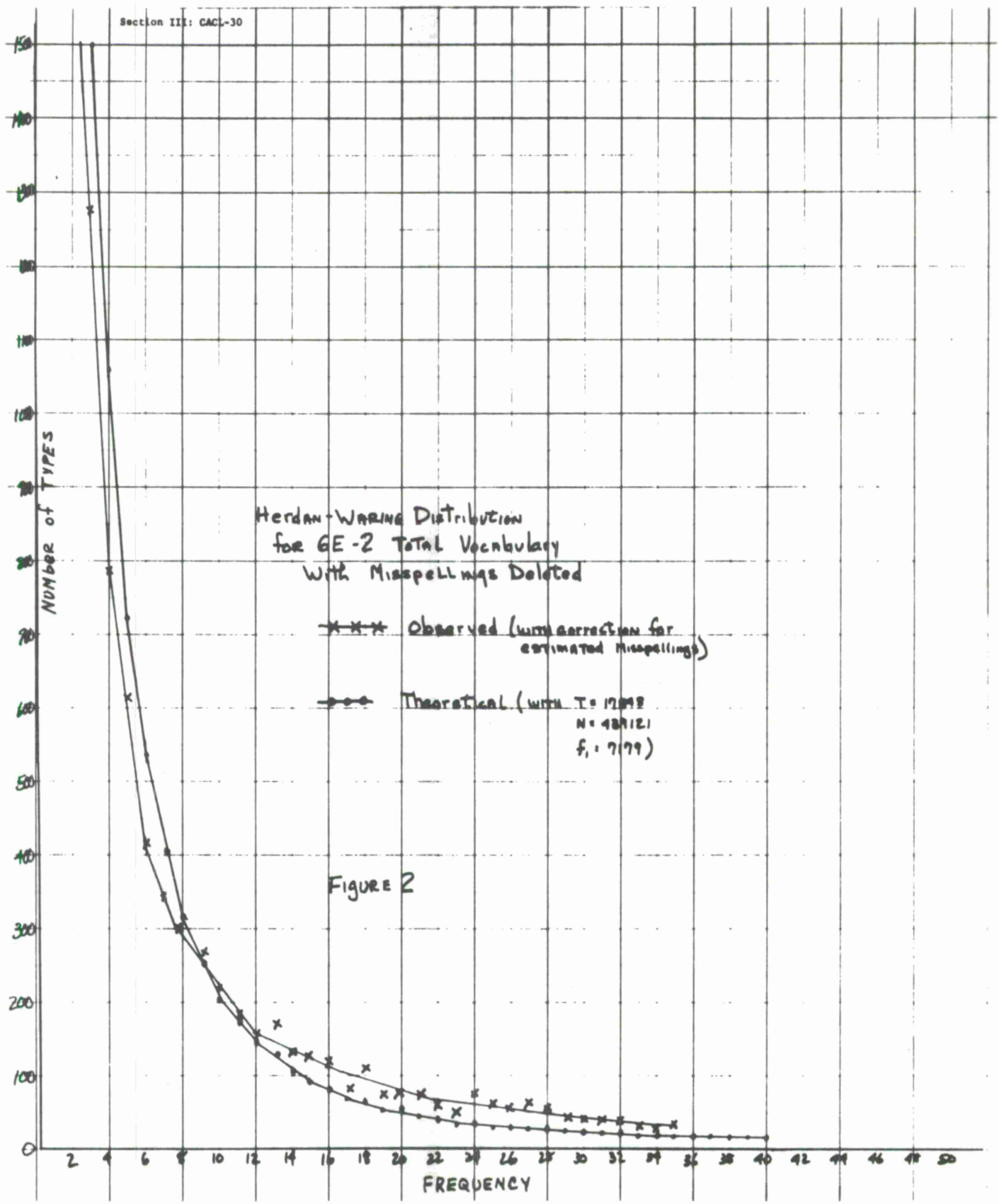

NUMBER of TYPES
FREQUENCY
Herdan-Waring Distribution
for GE-2 Total Vocabulary
With Misspellings Deleted
Observed (with correction for estimated misspellings)
Theoretical (with T= 17898
N= 488121
f1 = 7179)

FIGURE 2

Derived Rank Frequency Plots for Vocabularies Distributed in a Sample According to the Waring Distribution

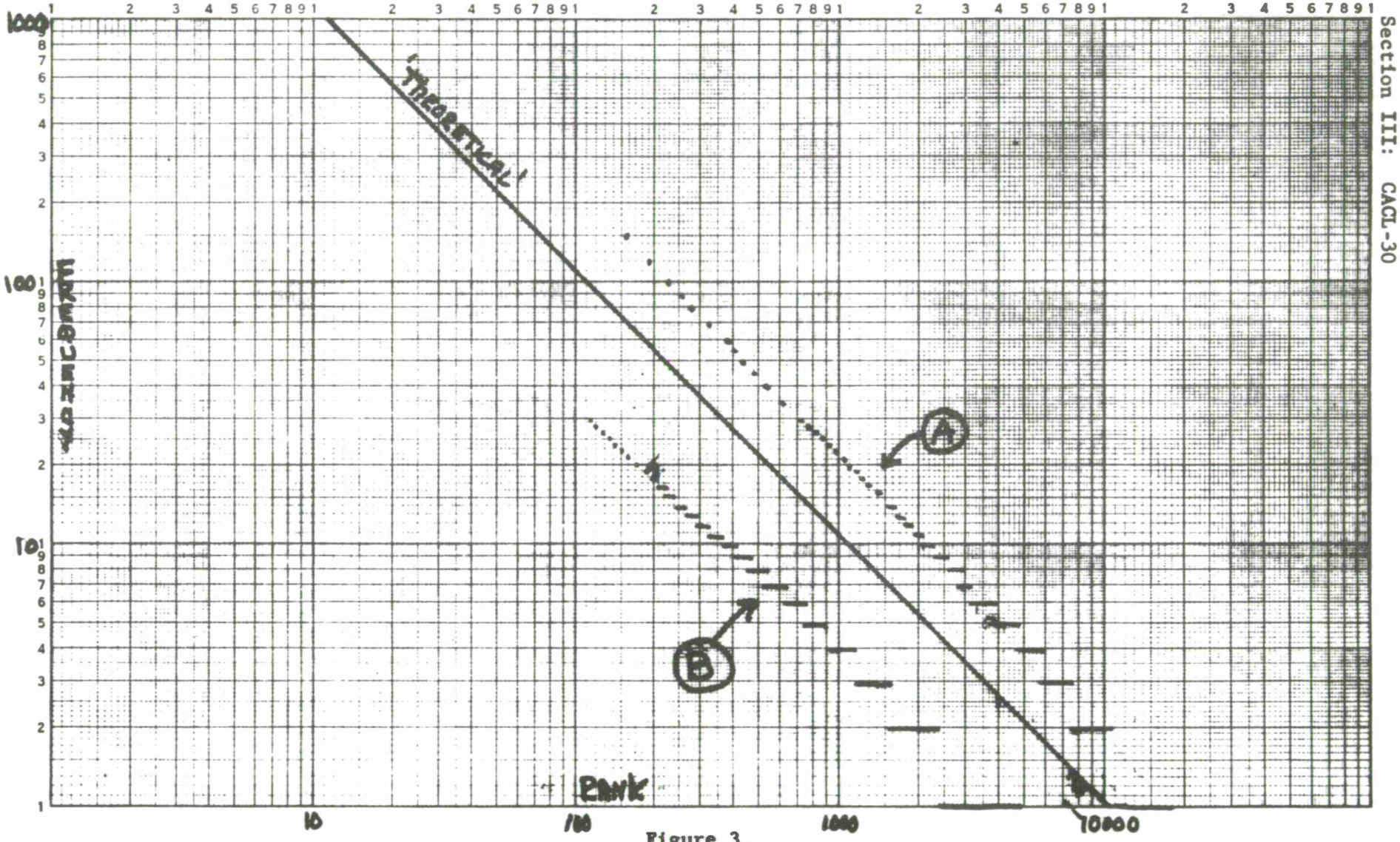


Figure 3.

Section III: Supplement to CACL-13

RANK VS. FREQUENCY PLOTS FOR 1, 2, 3, AND 4 WORD STRINGS IN G.E.-2 CORPUS *

This supplement contains the Rank-Frequency plots for the context strings counted in the processing of the G.E.-2 data. Several supplementary programs, prepared and run by J. Mehring, were used to obtain these counts.

The four plots are attached. The reader will note that the origin of the coordinate system for the first plot (single words) is different from that used for the other three figures.

Observations

1. The plot for one-word contexts (i.e., word forms) does not show the straight line with slope -1 appropriate to the Zipf curve. It has a distinct curvature which presumably can be attributed to the fragmented nature of the corpus and the heterogeneity of the subjects treated.

2. The plots for the context strings of various lengths all show rather straight lines with slope in the vicinity of $-\frac{1}{2}$. I think the straightness is interesting.

This supplement is being distributed for general interest. So far as I know, these are the first Zipf plots of context strings ever prepared, so providing a good explanation of these data would be more than sheer re-creation.

Reference

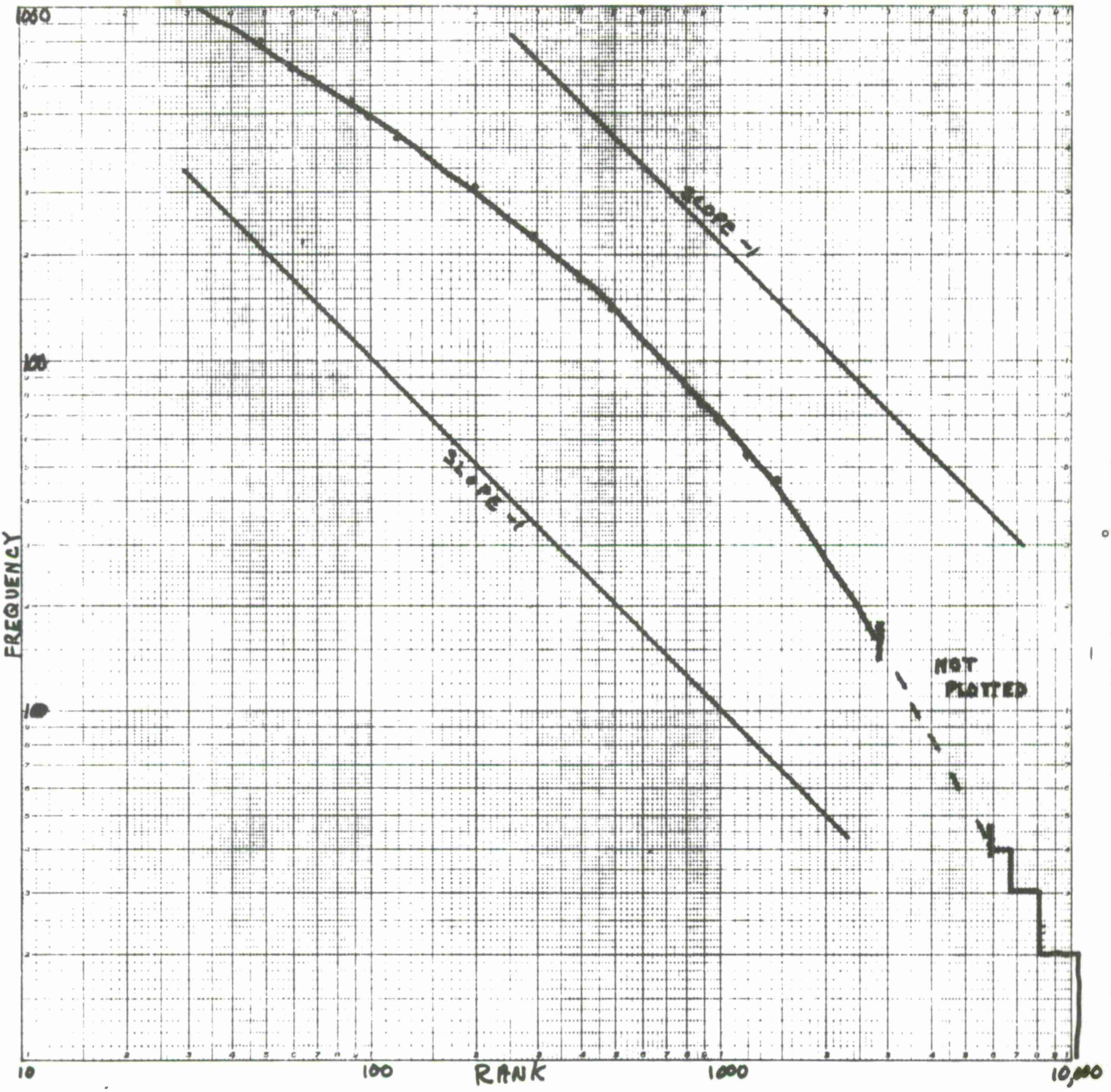
Zipf, G. K., Human Behavior and the Principle of Least Effort, Addison-Wesley, Cambridge, 1949.

---

* Issued on May 5, 1965 to a limited distribution as Supplement to Technical Note CACL-13 by Paul E. Jones, Jr.

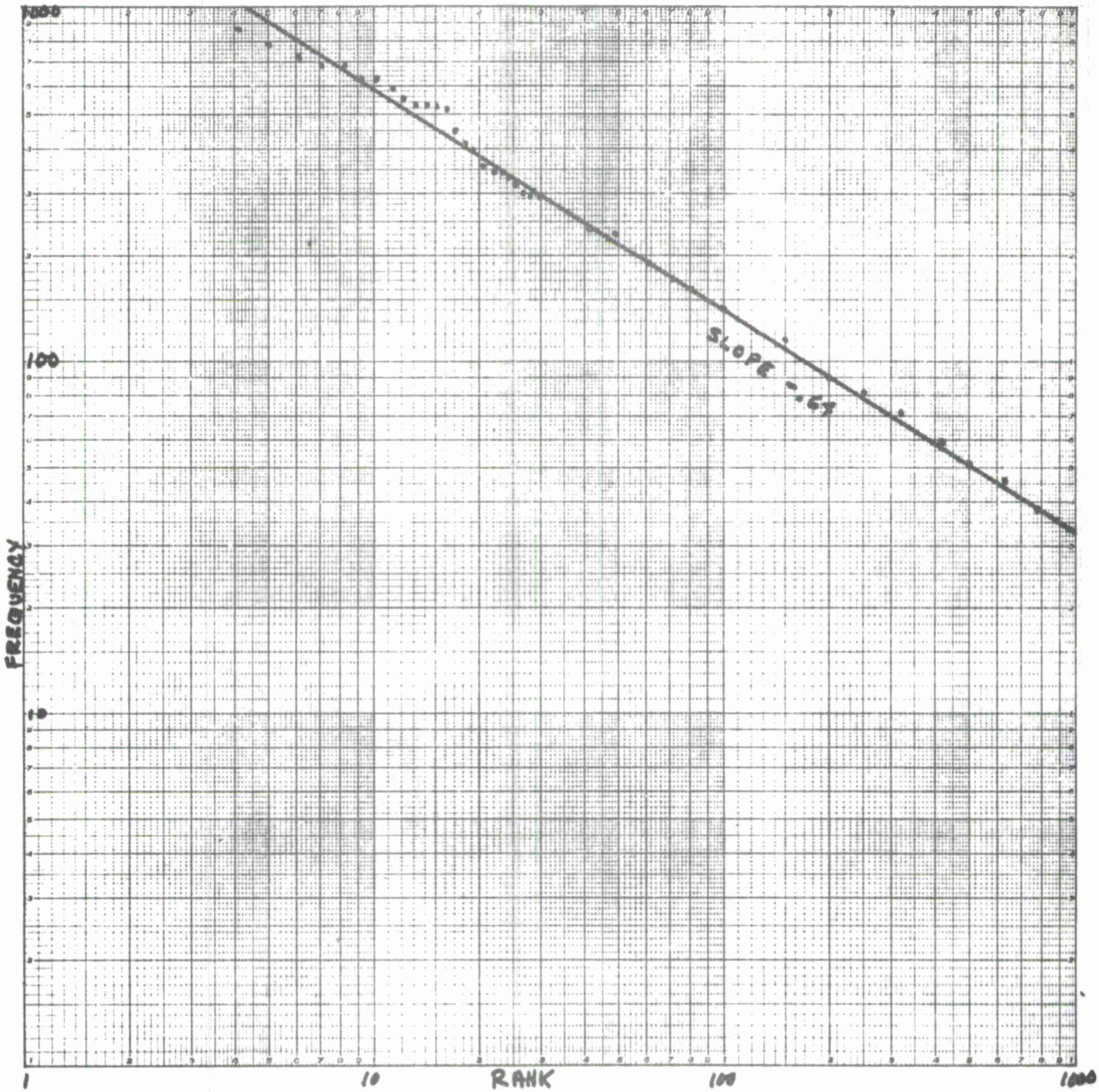
Section III: Supplement to CACL-13

Rank-Frequency Curve for Types Discovered in GE Collection
of ~ 446,000 Running Words in 10,287 Fragments



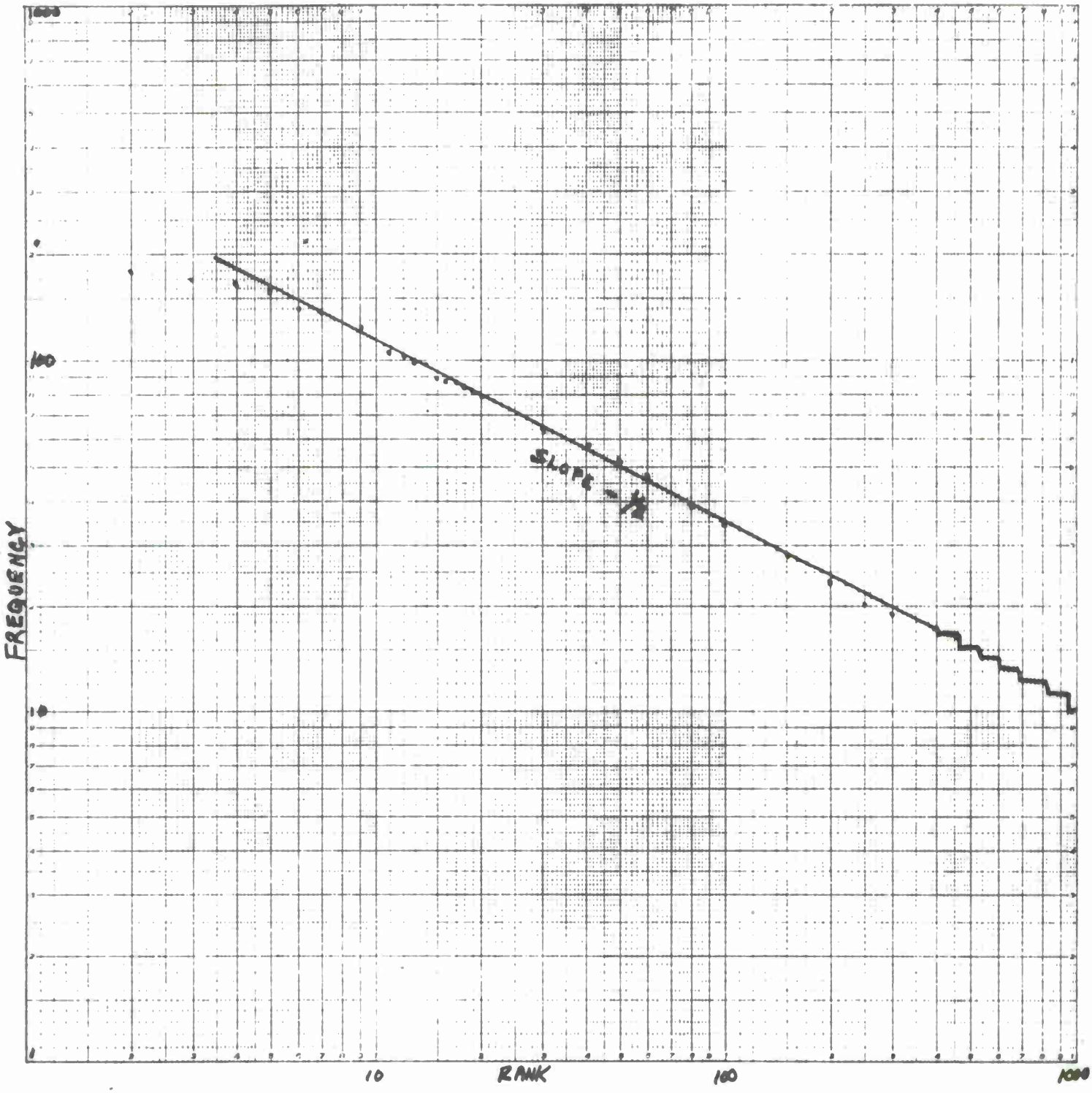
Section III: Supplement to CACL-13

Rank-Frequency Curve for Two-Word Contexts in GE-2 Collection
of ~446,000 Running Words in 10,287 Fragments



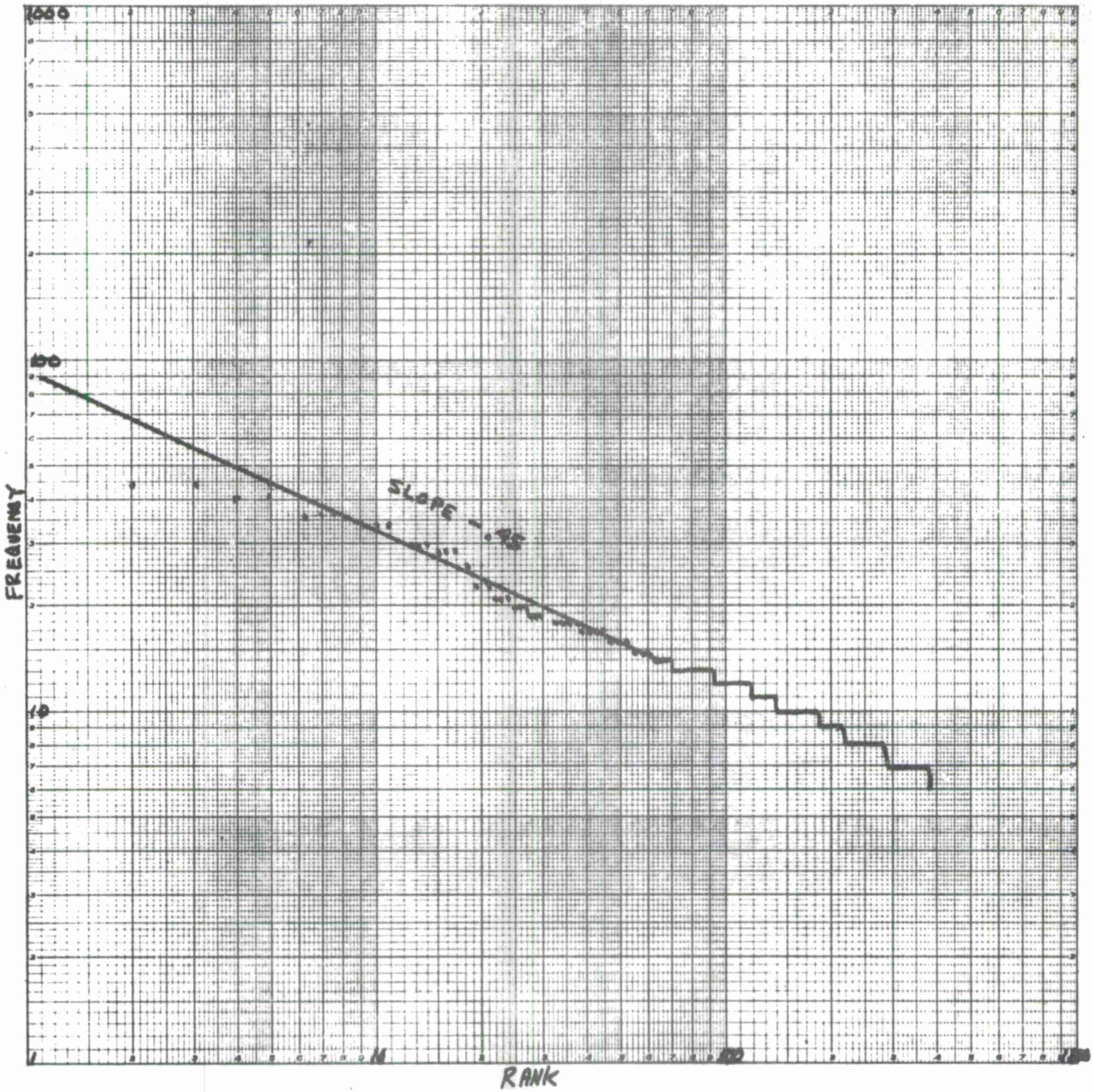
Section III: Supplement to CACL-13

Rank-Frequency Curve for Three-Word Contexts in GE-2
Collection of ~ 446,000 Running Words in 10,287 Fragments



Section III: Supplement to CACL-13

Rank-Frequency Curve for Four-Word Contexts in GE-2
Collection of ~446,000 Running Words in 10,287 Fragments



SECTION III: CACL-29

ZIPF CURVES FOR GE AND NASA INDEXING VOCABULARIES*

We have repeatedly observed that the large manual indexing vocabularies we have studied do not produce a straight-line Zipf curve. They characteristically are bowed. This note records the rank-frequency plots for two manual indexing vocabularies, the GE-0 vocabulary and NASA's.

A. Zipf Curve for GE-0 Manual Indexing Vocabulary

Figure 1 accurately shows the rank-frequency curve for the frequent GE terms, where the frequency plotted is a term's total postings in the 69,668-document GE-0 collection. Figures were obtained from the vocabulary listing dated November, 1962.

1. Data

All terms with frequency $>$ 100 are plotted and most, but not all, of the lower frequency terms are also plotted. This is because we had only a partial deck of terms (with frequencies) in keypunched form. To complete the deck, the missing terms were identified, their frequencies looked up, and a new card containing the missing frequency was prepared (for frequencies $>$ 100).

A total of 1,645 terms with frequency $<$ 100 are omitted from the plot. The dashed line shows how they would probably fit in. (Because there are 4,824 terms in the vocabulary, the curve has to reach frequency 1 at rank 4824.)

B. Zipf Curves for the NASA Vocabulary

Figure 2 shows the usual rank-frequency plot of the whole 18,292 NASA indexing vocabulary, reflecting postings to the subcollection of about 100,000 documents we have studied. The plot shows 10,083 postings to the term of rank 1 and drops to 0 postings at about rank 16,000. This curve is definitely non-Zipfian, exhibiting a strong bow upwards.

C. The Division of the NASA Vocabulary into Two Separate Vocabularies

Due to the high density of multiple word terms (MWT) near the low frequency end of the NASA vocabulary list, it was decided to actually tabulate the distribution of the MWT's over the whole vocabulary. The number of MWT's on each page of the NASA's frequency-ordered index term dictionary were sampled and the proportions were used to obtain separate rank vs. frequency plots for the MWT subvocabulary and the Single Word Term subvocabulary as shown in Figures 3 and 4 respectively. Again, each of these curves is seen to be decidedly non-Zipfian.

* Issued on May 6, 1966 to a limited distribution by Peter R. Bono and Paul E. Jones, Jr. as Technical Note CACL-29.

Zipf Curve for GE-0 Indexing Vocabulary

(Accurate for f ≥ 100; Approximate for f < 100)

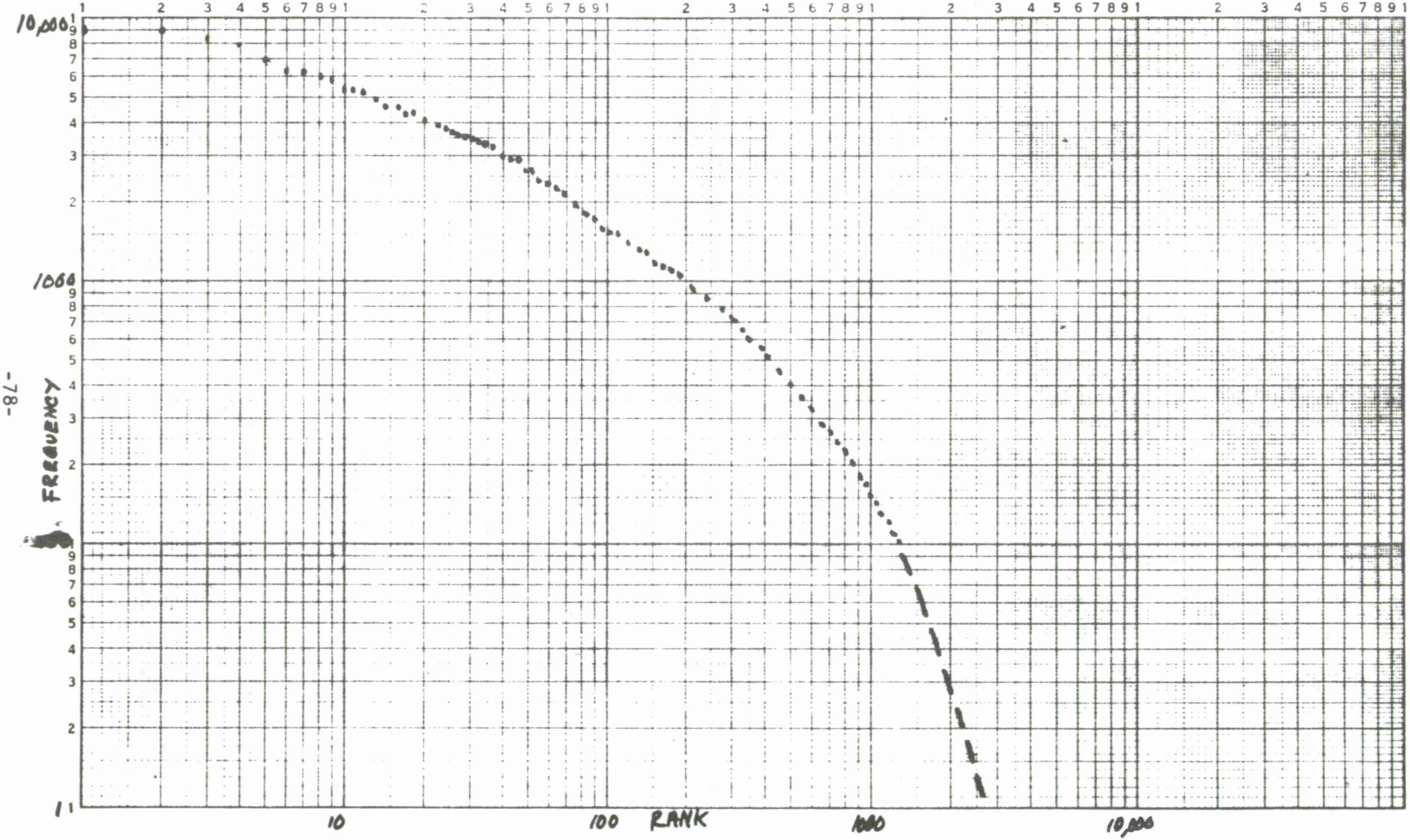
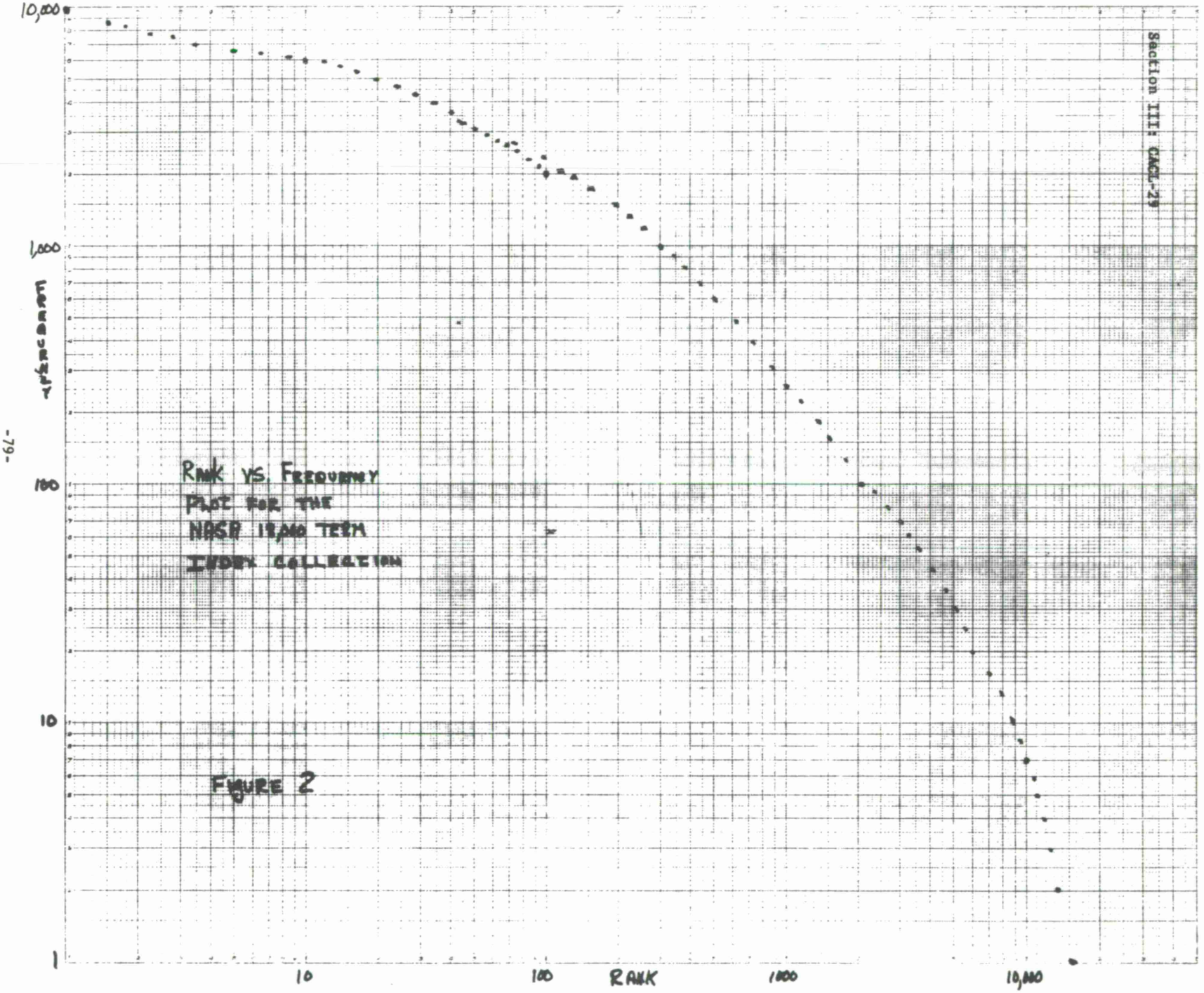


Figure 1



RANK VS. FREQUENCY PLOT FOR THE NASA 18,000 TERM INDEX COLLECTION

FIGURE 2

Rank vs. Frequency for the Multiple-Word Terms only in the NASA 18 K Index Collection

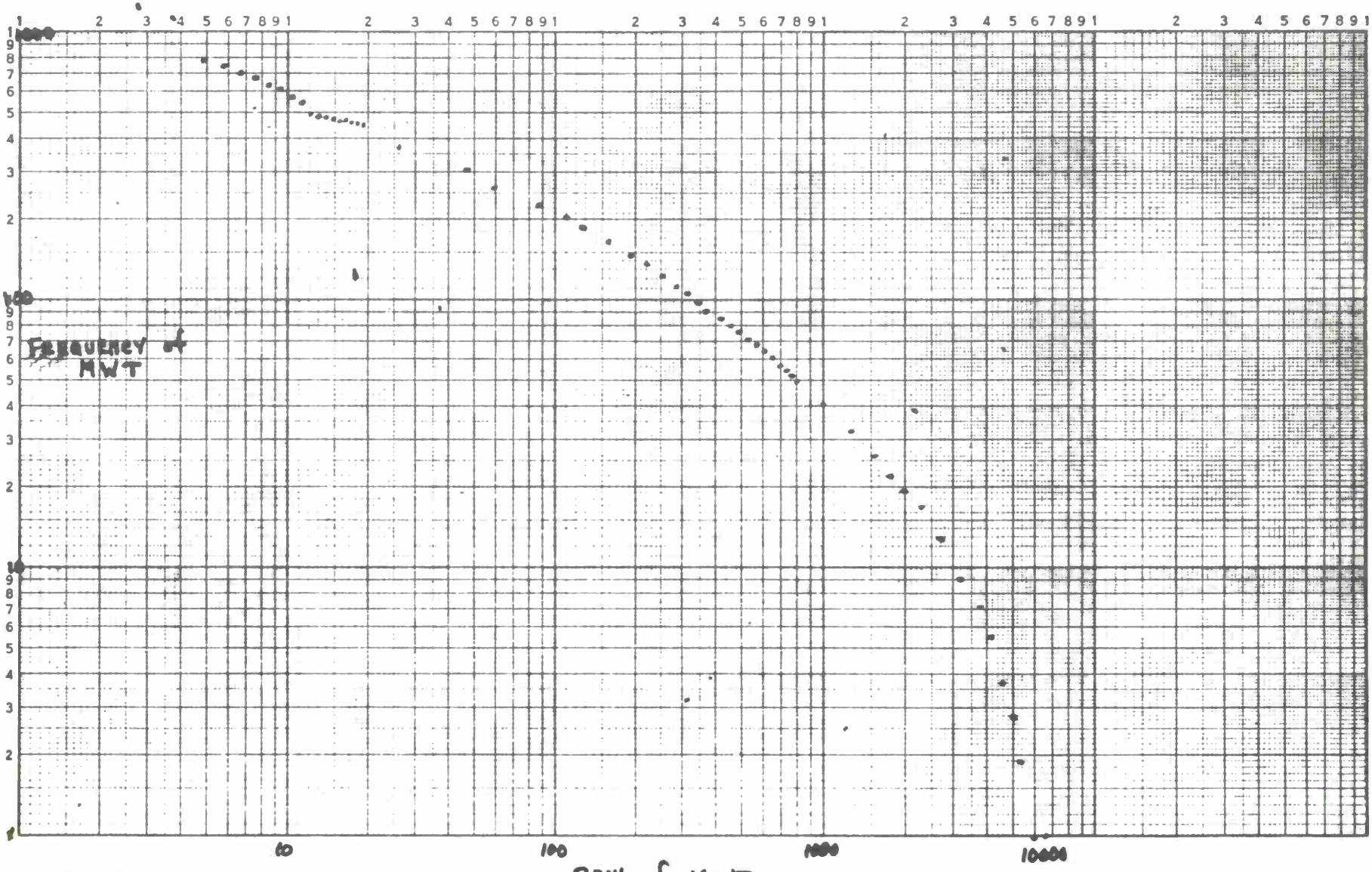
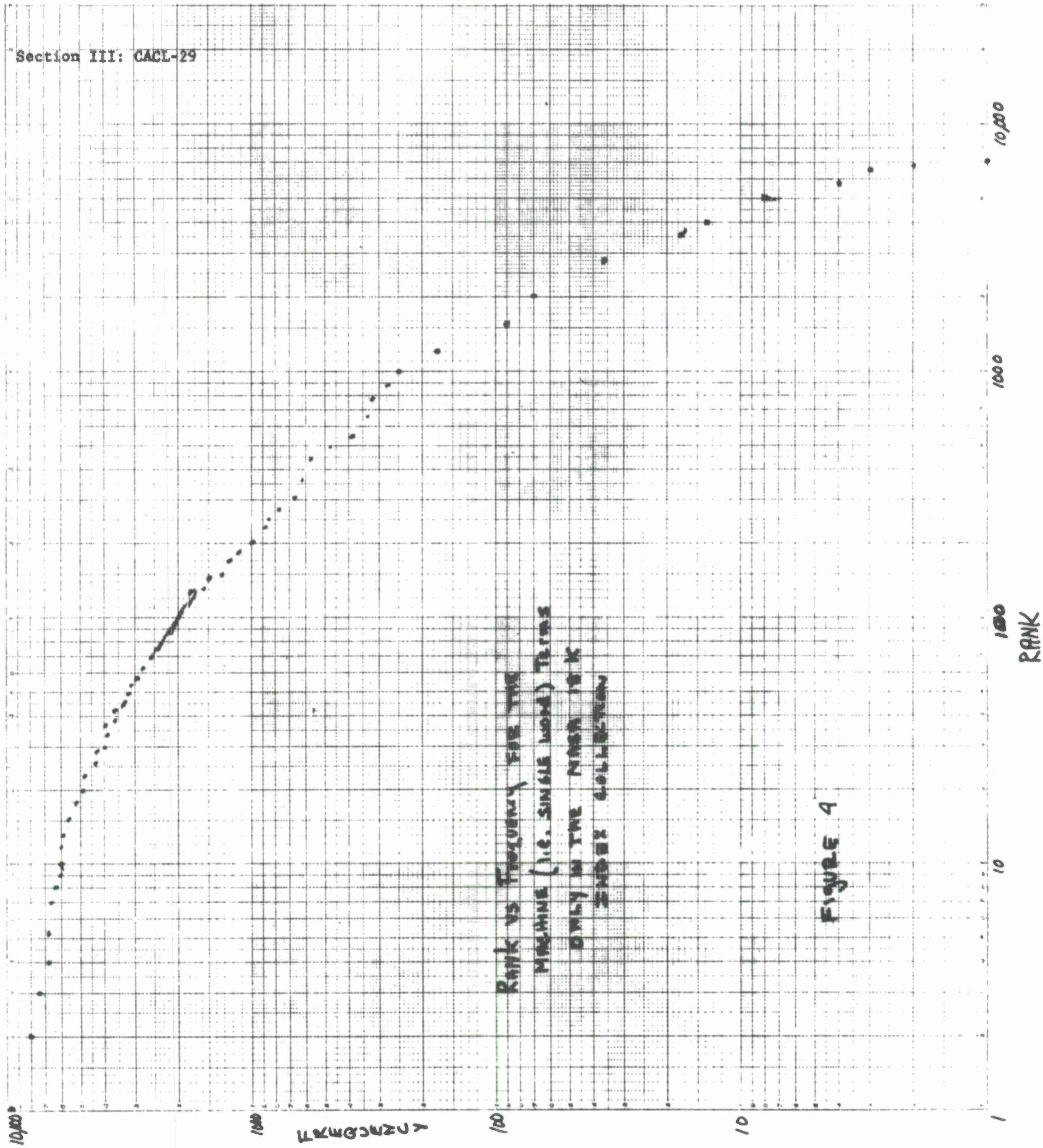


Figure 3


RANK vs FREQUENCY FOR THE MACHINE (i.e. SINGLE WORD) TERMS ONLY IN THE NASA 18K INDEX COLLECTION
FIGURE 4
RANK
FREQUENCY
10,000
1000
100
10
1

SECTION IV

STUDIES OF CONTENT BEARING UNITS IN TEXT

Selecting Content Bearing Units*

In order to understand the work on overlap described in the next paper, it is necessary to present a few details concerning the processing of the data, in addition to that in CACL-13 in Section I. The GE-2 abstract collection was processed to reveal all two-word strings, the frequencies of the 2 terms involved, and the frequency of the string itself. This processing has been described in Section I. Of those two-word strings, a content-bearing unit was defined as a string AB which occurred fab times, consisting of two terms with frequencies fa and fb, if the following conditions were met:

1) $fab \geq 3$

2) $fa \leq 2040$

3) $fb \leq 2040$

4) $Cab \geq 20$

where

$$Cab = \frac{fab \times 450{,}000}{fa \times fb}$$

In order to show the exact nature of this data and as possible sample data for those who wish to test other methodologies we present a portion of the two-word strings derived from the GE data base. The listing below (Figure 1) shows all recurrent pairs whose first word begins with the letter "O". The pair itself, the frequencies of the two words involved, and the frequency of the pair are respectively listed. In addition, the cohesion value (Cab) of those pairs which meet conditions (1) - (4) is given.

FIGURE 1

Word A	Word B	f_a	f_b	fab	Cab
O16	T	2000002	000085	00000000000000002	
O2	H2	2000016	000033	00000000000000003	2556
O2	N2	2000016	000016	00000000000000002	
O2	AND	2000016	016622	00000000000000003	
OBEYING	A	2000004	006232	00000000000000002	
OBJECTIVE	IN	2000035	008879	00000000000000003	
OBJECTIVE	IS	2000035	001704	00000000000000003	22

* By Robert M. Curtice. Not previously issued.

Table 1 (Continued)

OBTAINED	FROM	2000545001807
OBTAINED	IF	2000545000113
OBTAINED	IN	2000545008879
OBTAINED	IS	2000545001704
OBTAINED	ONLY	2000545000241
OBTAINED	ON	2000545004634
OBTAINED	OVER	2000545000424
OBTAINED	TO	2000545008045
OBTAINED	UNDER	2000545000596
OBTAINED	WHEN	2000545000303
OBTAINED	WHICH	2000545001180
OBTAINED	WITH	2000545003844
OBTAINED	THROUGH	2000545000402
OBTAINED	USING	2000545000572
OBTAINING	ADEQUATE	2000064000043
OBTAINING	DATA	2000064000943
OBTAINING	HOT	2000064000236
OBTAINING	AN	2000064001474
OBTAINING	A	2000064006232
OBTAINING	THE	2000064008911
OBTAINING	SOLUTIONS	2000064000390
OBTAINING	STEELS	2000064000311
OBTAINING	STRESS	2000064000890
OBTAIN	ANALYTICAL	2000106000181
OBTAIN	DATA	2000106000943
OBTAIN	DESIGN	2000106001173
OBTAIN	GENERAL	2000106000387
OBTAIN	INFORMATION	2000106000224
OBTAIN	INSIGHT	2000106000008
OBTAIN	MAXIMUM	2000106000202
OBTAIN	NUMERICAL	2000106000189
OBTAIN	QUALITATIVE	2000106000032
OBTAIN	AN	2000106001474
OBTAIN	A	2000106006232
OBTAIN	THE	2000106008911
OBTAIN	SATISFACTORY	2000106000083
OBTAIN	SOLUTIONS	2000106000390
OBTENUS	PAR	2000002000019
OCCUPIED	BY	2000002002759
OCCURENCE	OF	2000004030170
OCCURRED	AT	2000024002770
OCCURRED	BY	2000024002759
OCCURRED	WHEN	2000024000303
OCCURRED	WITH	2000024003844
OCCURRENCE	AND	2000010016622
OCCURRENCE	OF	2000010030170
OCCURRING	DURING	2000019000286
OCCURRING	IN	2000019008879
OCCUR	DISCUSSED	2000046000497
OCCUR	M	2000046007364
OCCUR	AND	2000046016622
OCCUR	AT	2000046002770

Table 1 (Continued)

OCCUR	DURING	20000460002860000000000000000002(
OCCUR	IN	20000460088790000000000000000013(
OCCUR	WHEN	20000460003030000000000000000005+161
OCCURS	FIRST	20000540001750000000000000000002(
OCCURS	AND	20000540166220000000000000000003(
OCCURS	AS	20000540016890000000000000000002(
OCCURS	AT	20000540027700000000000000000005(
OCCURS	FOR	20000540074060000000000000000004(
OCCURS	IN	20000540088790000000000000000012(
OCCURS	WHEN	20000540003030000000000000000003(
OCCURS	WITH	20000540038440000000000000000002(
OCTENE	1	20000030006320000000000000000003+712
OCTOBER	1952	20000110000170000000000000000002(
OCTOBER	1	20000110006320000000000000000002(
OCT	1954	20000110000140000000000000000002(
OCT	1955	20000110000140000000000000000002(
OCT	1	20000110006320000000000000000002(
OFFERED	NO	20000080003720000000000000000002(
OFFICE	OF	20000040301700000000000000000003(
OFF	AIRPLANE	20000880000840000000000000000002(
OFF	DESIGN	20000880011730000000000000000010+23
OFF	LIMITS	20000880000920000000000000000002(
OFF	M	20000880073640000000000000000002(
OFF	AND	20000880166220000000000000000013(
OFF	SIZE	20000880020700000000000000000002(
OFF	THRUST	20000880004180000000000000000005+61
OFF	VALVE	20000880000740000000000000000002(
OFFSET	DIFFUSERS	20000090000300000000000000000002(
OFFSET	YIELD	20000090001390000000000000000004+488
OF.005	PCT	20000020002410000000000000000002(
OFTEN	FOUND	20000090002630000000000000000002(
OGIVE	CYLINDER	20000050002200000000000000000003+1227
OHIO	IT	20000080002600000000000000000002(
OHIO	STATE	20000080002690000000000000000002(
OHMS	LAW	20000030000590000000000000000002(
OIL	+	20001200004260000000000000000002(
OIL	ASH	20001200000190000000000000000003+750
OIL	BURNING	20001200001420000000000000000004+105
OIL	COLUMN	20001200000430000000000000000002(
OIL	DEVELOPMENT	20001200008910000000000000000002(
OIL	FILM	20001200001530000000000000000008+196
OIL	FLOW	20001200021840000000000000000004(
OIL	HOSE	20001200000100000000000000000002(
OIL	M	20001200073640000000000000000002(
OIL	QPR	20001200000620000000000000000002(
OIL	RESULTS	20001200011160000000000000000002(
OIL	AND	20001200166220000000000000000008(
OIL	AT	20001200027700000000000000000002(
OIL	IN	20001200088790000000000000000003(
OIL		20001200000080000000000000000002(

Table 1 (Continued)

OBJECTIVE	OF	20000350301700000000000000000011
OBJECTIVE	TO	20000350080450000000000000000003
OBJECTIVE	WAS	20000350003980000000000000000005 161
OBJECTIVES	AND	20000200166220000000000000000004
OBJECTIVES	OF	20000200301700000000000000000010
OBJECT	RETENTION	20000470000600000000000000000002
OBJECT	OF	20000470301700000000000000000029
OBJECT	TO	20000470080450000000000000000005
OBJECT	WAS	20000470003980000000000000000002
OBJECTS	PRODUCED	20000080001910000000000000000002
OBJECTS	OF	20000080301700000000000000000002
OBLATENESS	AND	20000040166220000000000000000002
OBLATE	SPHEROIDAL	20000030000020000000000000000002
OBLIQUE	COORDINATES	20000200004700000000000000000002
OBLIQUE	SHOCK	20000200005970000000000000000012 452
OBRABOTKA	METALLOV	20000030000110000000000000000003 40909
OBSCURED	BY	20000020027590000000000000000002
OBSERVATION	OF	20000240301700000000000000000015
OBSERVATIONSMADE		20000690005180000000000000000003 37
OBSERVATIONS	AND	20000690166220000000000000000002
OBSERVATIONS	OF	20000690301700000000000000000013
OBSERVATIONS	ON	20000690046340000000000000000026
OBSERVED	CHANGES	20001010001540000000000000000002
OBSERVED	MEASURED	20001010001990000000000000000002
OBSERVED	M	20001010073640000000000000000007
OBSERVED	AND	20001010166220000000000000000005
OBSERVED	BY	20001010027590000000000000000002
OBSERVED	DURING	20001010002860000000000000000002
OBSERVED	FOR	20001010074060000000000000000004
OBSERVED	IN	20001010088790000000000000000016
OBSERVED	THAT	20001010000540000000000000000002
OBSERVED	TO	20001010080450000000000000000002
OBSERVED	UNDER	20001010005960000000000000000002
OBSERVED	WITH	20001010038440000000000000000004
OBSERVED	STRENGTH	20001010006890000000000000000002
OBSERVED	VALUES	20001010002400000000000000000003 55
OBTAINABLE	NOISE	20000600002050000000000000000002
OBTAINED	AF33	20005450000940000000000000000002
OBTAINED	DIRECTLY	20005450000480000000000000000002
OBTAINED	EFFECTS	20005450008310000000000000000003
OBTAINED	EXPERIMENTAL	20005450008810000000000000000002
OBTAINED	M	20005450073640000000000000000011
OBTAINED	PREVIOUSLY	20005450000590000000000000000002
OBTAINED	AND	20005450166220000000000000000011
OBTAINED	AS	20005450016890000000000000000007
OBTAINED	AT	20005450027700000000000000000018
OBTAINED	BETWEEN	20005450006760000000000000000003
OBTAINED	BOTH	20005450002720000000000000000002
OBTAINED	BY	20005450027590000000000000000082
OBTAINED	DURING	20005450002860000000000000000007 20
OBTAINED	FOR	20005450074060000000000000000087

Table 1 (Continued)

OIL	TO	2000120008045000000000000000003(
OIL	SOAKED	2000120000002000000000000000002(
OIL	SUPPLY	2000120000076000000000000000002(
OIL	SYSTEM	2000120000916000000000000000002(
OIL	THICKENER	2000120000003000000000000000002(
OIL	TRANSFER	2000120000791000000000000000002(
OIL	VISCOSITY	2000120000117000000000000000002(
OILS	AND	2000062016622000000000000000009(
OILS	AT	2000062002770000000000000000002(
OILS	HAVING	2000062000203000000000000000002(
OILS	IN	2000062008879000000000000000005(
OILS	ON	2000062004634000000000000000002(
OKLAHOMA	CITY	2000002000005000000000000000002(
OLEIC	ACID	2000002000098000000000000000002(
OLEOPHOBIC	FILMS	2000002000073000000000000000002(
OMEGA	PHASE	2000007000340000000000000000002(
ONDE	DE	2000002000163000000000000000002(
ONEDIMENSION	UNSTEADY	2000008000072000000000000000002(
ONERA	M	2000010007364000000000000000002(
ONR	SPONSORED	2000015000028000000000000000010 10719
ONR	SUPPORTED	2000015000094000000000000000002(
ONSET	OF	2000011030170000000000000000011(
ONS	OF	2000004030170000000000000000002(
OPENING	IN	2000014008879000000000000000003(
OPENING	OF	2000014030170000000000000000002(
OPENINGS	IN	2000006008879000000000000000005(
OPEN	CYCLE	2000055000126000000000000000003 194
OPEN	ENDED	2000055000008000000000000000002(
OPEN	END	2000055000094000000000000000003 261
OPEN	IMPELLER	2000055000032000000000000000002(
OPEN	JET	2000055000618000000000000000002(
OPEN	LOOP	2000055000059000000000000000004 559
OPEN	M	2000055007364000000000000000003(
OPEN	NOSE	2000055000061000000000000000004 536
OPEN	AND	2000055016622000000000000000004(
OPERAM	TION	2000003000037000000000000000002(
OPERATED	AS	2000034001689000000000000000003 28
OPERATED	AT	2000034002770000000000000000013(
OPERATED	FROM	2000034001807000000000000000002(
OPERATED	IN	2000034008879000000000000000003(
OPERATED	WITH	2000034003844000000000000000004(
OPERATED	SERVO	2000034000049000000000000000002(
OPERATE	AND	2000037016622000000000000000002(
OPERATE	AT	2000037002770000000000000000006(
OPERATE	IN	2000037008879000000000000000006(
OPERATE	ON	2000037004634000000000000000003(
OPERATE	WITHIN	2000037000138000000000000000002(
OPERATES	IN	2000008008879000000000000000002(
OPERATING	CHARACTERIST	2000229000833000000000000000013 80
OPERATING	CONDITION	2000229000123000000000000000003 47
OPERATING	CONDITIONS	2000229000703000000000000000029 81
OPERATING	COST	2000229000122000000000000000005 80

Table 1 (Continued)

OPERATING	COSTS	2000229000063
OPERATING	DATA	2000229000943
OPERATING	EFFICIENCY	2000229000197
OPERATING	ENVIRONMENT	2000229000065
OPERATING	EXPERIENCE	2000229000054
OPERATING	LIMITS	2000229000092
OPERATING	MECHANISM	2000229000191
OPERATING	M	2000229007364
OPERATING	NEAR	2000229000130
OPERATING	PRINCIPLES	2000229000119
OPERATING	RANGE	2000229000582
OPERATING	AND	2000229016622
OPERATING	A	2000229006232
OPERATING	AT	2000229002770
OPERATING	IN	2000229008879
OPERATING	ON	2000229004634
OPERATING	OVER	2000229000424
OPERATING	UNDER	2000229000596
OPERATING	WITH	2000229003844
OPERATING	SAME	2000229000119
OPERATING	SPEED	2000229000488
OPERATING	SUPERSONICAL	2000229000002
OPERATING	TEMPERATURE	2000229002033
OPERATING	TEMPERATURES	2000229000726
OPERATING	VARIABLES	2000229000142
OPERATIN	IN	2000003008879
OPERATIONAL	ASPECTS	2000102000120
OPERATIONAL	CALCULUS	2000102000013
OPERATIONAL	CHARACTERIST	2000102000833
OPERATIONAL	CONDITIONS	2000102000703
OPERATIONAL	CRITERIA	2000102000107
OPERATIONAL	DIFFICULTIES	2000102000032
OPERATIONAL	EXPERIENCE	2000102000054
OPERATIONAL	INFORMATION	2000102000224
OPERATIONAL	INTEGRITY	2000102000006
OPERATIONAL	LIMITS	2000102000092
OPERATIONAL	PERFORMANCE	2000102000774
OPERATIONAL	PRINCIPLES	2000102000119
OPERATIONAL	PROBLEMS	2000102000619
OPERATIONAL	REQUIREMENTS	2000102000282
OPERATIONAL	AND	2000102016622
OPERATIONAL	SPEEDS	2000102000221
OPERATIONAL	STATUS	2000102000059
OPERATIONAL	TEMPERATURES	2000102000726
OPERATION	DATA	2000269000943
OPERATION	DESCRIBED	2000269000396
OPERATION	DISCUSSED	2000269000497
OPERATION	INCLUDING	2000269000162
OPERATION	INVESTIGATIO	2000269001078
OPERATION	M	2000269007364
OPERATION	AND	2000269016622
OPERATION	ARE	2000269000996
OPERATION	A	2000269006232

Table 1 (Continued)

```
OPERATION     AS          2000269001689000000000000000002(
OPERATION     AT          2000269002770000000000000000016(
OPERATION     IN          2000269008879000000000000000006(
OPERATION     IS          2000269001704000000000000000003(
OPERATION     OF          2000269030170000000000000000073(
OPERATION     ON          2000269004634000000000000000006(

OPERATION     SOME        2000269000545000000000000000003(
OPERATION     TO          2000269008045000000000000000003(
OPERATION     WERE        2000269000355000000000000000002(
OPERATION     WITH        2000269003844000000000000000005(
OPERATIONS    INVOLVED    2000084000080000000000000000002(
OPERATIONS    RESEARCH    2000084000575000000000000000010(193
OPERATIONS    AND         2000084016622000000000000000004(
OPERATIONS    FOR         2000084007406000000000000000002(
OPERATIONS    IN          2000084008879000000000000000003(
OPERATIONS    OF          2000084030170000000000000000006(
OPERATIONS    ON          2000084004634000000000000000003(
OPERATIONS    OVER        2000084000424000000000000000002(

OPERATIONS    TO          2000084008045000000000000000002(
OPERATIVE     AT          2000010002770000000000000000002(
OPERATIVE     TO          2000010008045000000000000000002(
OPERATOR      OF          2000018030170000000000000000002(
OPERATORS     WITH        2000011003844000000000000000002(
OPPORTUNITY   FOR         2000003007406000000000000000002(
OPPOSITE      DIRECTIONS  2000013000016000000000000000002(

OPPOSITE      EDGES       2000013000059000000000000000004(2346
OPTICALLY     ACTIVE      2000004000021000000000000000002(
OPTICAL       ABSORPTION  2000074000066000000000000000004(368
OPTICAL       MASER       2000074000003000000000000000002(
OPTICAL       MASERS      2000074000003000000000000000002(
OPTICAL       MEASUREMENTS 2000074000437000000000000000002(
OPTICAL       MECHANICAL  2000074000417000000000000000002(
OPTICAL       METHODS     2000074000755000000000000000002(
OPTICAL       MICROSCOPY  2000074000019000000000000000002(
OPTICAL       PUMPING     2000074000037000000000000000002(
OPTICAL       PYROMETER   2000074000016000000000000000003(1140
OPTICAL       PYROMETRY   2000074000010000000000000000002(
OPTICAL        AND        2000074016622000000000000000007(
OPTICAL       STUDIES     2000074000458000000000000000002(
OPTICAL       SYSTEM      2000074000916000000000000000005(33
OPTICAL       TECHNIQUES  2000074000425000000000000000002(
OPTICAL       TEMPERATURE 2000074002033000000000000000002(
OPTIMALIZINGCRUISE        2000004000018000000000000000002(
OPTIMAL       CONTROL     2000019000703000000000000000004(139
OPTIMAL       PERFORMANCE 2000019000774000000000000000002(
OPTIMIZATIONPROBLEMS      2000029000619000000000000000002(
OPTIMIZATION FOR          2000029007406000000000000000003(
OPTIMIZATION OF           2000029030170000000000000000010(
OPTIMIZATIONSTUDY         2000029000862000000000000000003(54
OPTIMIZING    SPACE       2000011000453000000000000000002(
OPTIMUM       BURNING     2000178000142000000000000000002(
```

Table 1 (Continued)

OPTIMUM	CODING	2000178000010000000000000000002
OPTIMUM	COLOR	2000178000015000000000000000002
OPTIMUM	CONDITIONS	2000178000703000000000000000003
OPTIMUM	CONTROL	2000178000703000000000000000006 21
OPTIMUM	CROSS	2000178000169000000000000000002
OPTIMUM	DESIGN	2000178001173000000000000000008
OPTIMUM	DESIGNS	2000178000081000000000000000002
OPTIMUM	EXHAUST	2000178000150000000000000000002
OPTIMUM	FREE	2000178000416000000000000000002
OPTIMUM	GEOMETRY	2000178000086000000000000000002
OPTIMUM	HEAT	2000178001547000000000000000002
OPTIMUM	METHODS	2000178000755000000000000000004
OPTIMUM	NONLINEAR	2000178000141000000000000000002
OPTIMUM	NOZZLE	2000178000400000000000000000004 25
OPTIMUM	PERFORMANCE	2000178000774000000000000000007 22
OPTIMUM	POLICY	2000178000004000000000000000002
OPTIMUM	PRESSURES	2000178000197000000000000000002
OPTIMUM	PROPERTIES	2000178001301000000000000000002
OPTIMUM	PROPORTIONS	2000178000009000000000000000003 842
OPTIMUM	RESPONSE	2000178000182000000000000000003 41
OPTIMUM	RUPTURE	2000178000213000000000000000002
OPTIMUM	WHEN	2000178000303000000000000000002
OPTIMUM	SHAPE	2000178000131000000000000000003 57
OPTIMUM	SIZE	2000178000207000000000000000002
OPTIMUM	TECHNIQUE	2000178000261000000000000000002
OPTIMUM	THRUST	2000178000418000000000000000002
OPTIMUM	TRANSIENT	2000178000157000000000000000002
OPTIMUM	VALUES	2000178000240000000000000000002
ORBITAL	RENDEZVOUS	2000019000006000000000000000002
ORBIT	PLANE	2000045000250000000000000000002
ORBIT	AND	2000045016622000000000000000002
ORBIT	OF	2000045030170000000000000000003
ORBIT	OR	2000045000906000000000000000002
ORBIT	TO	2000045008045000000000000000002
ORBIT	WITH	2000045003844000000000000000003
ORBITS	ABOUT	2000021000204000000000000000002
ORBITS	IN	2000021008879000000000000000002
ORBITS	WITH	2000021003844000000000000000002
ORDER	APPROXIMATIO	2000198000112000000000000000004 81
ORDER	BOUNDARY	2000198000901000000000000000002
ORDER	DIFFERENTIAL	2000198000213000000000000000005 63
ORDER	DISORDER	2000198000002000000000000000002
ORDER	M	2000198007364000000000000000002
ORDER	NONLINEAR	2000193000141000000000000000003 48
ORDER	PARTIAL	2000198000097000000000000000002
ORDER	PROCESS	2000198000349000000000000000002
ORDER	AND	2000198016622000000000000000002
ORDER	FOR	2000198007406000000000000000003
ORDER	IN	2000198008879000000000000000003
ORDER	OF	2000198030170000000000000000034
ORDER	ONE	2000198000381000000000000000002
ORDER		2000198000008000000000000000004
ORDER	THAT	2000198000954000000000000000040

Table 1 (Continued)

```
ORDER          TO                2000198008045000000000000000048(
ORDER          WITH              2000198003844000000000000000002(
ORDER          SLENDER           2000198000068000000000000000002(
ORDER          SOLUTIONS         2000198000390000000000000000002(
ORDER          SUBSONIC          2000198000133000000000000000002(
ORDER          SYSTEMS           2000198000757000000000000000004(
ORDER          TERMS             2000198000224000000000000000003(30
ORDER          THEORY            2000198000808000000000000000005(
ORDER          VELOCITY          2000198000541000000000000000002(
ORDER          WAVE              2000198000371000000000000000002(
ORDERS          OF               2000008030170000000000000000005(
ORDINARY       DIFFERENTIAL      2000037000213000000000000000011(438
ORDINARY       PLANE             2000037000250000000000000000002(
ORDINARY       TEMPERATURE       2000037002033000000000000000002(
ORDINATE       SYSTEMS           2000030000757000000000000000002(
ORD            2059              2000042000011000000000000000011(10714
ORD            23004             2000042000002000000000000000002(
ORD            2787              2000042000002000000000000000002(
ORD            4862              2000042000002000000000000000002(
ORD            9                 2000042000049000000000000000002(
ORDVAC          AND              2000003016622000000000000000002(

ORGANIC        BONDING           2000063000071000000000000000002(
ORGANIC        COATINGS          2000063000275000000000000000005(129
ORGANIC        COMPOUNDS         2000063000153000000000000000016(796
ORGANIC        DERIVATIVES       2000063000052000000000000000002(
ORGANIC        FLUORINE          2000063000050000000000000000002(
ORGANIC        LIQUIDS           2000063000075000000000000000003(285
ORGANIC        LUBRICANTS        2000063000156000000000000000002(
ORGANIC        MATERIALS         2000063001020000000000000000005(35
ORGANIZATION AND                 2000026016622000000000000000002(
ORGANIZATION OF                  2000026030170000000000000000003(
ORGANOMETALLCOMPOUNDS            2000004000153000000000000000002(
ORGANO         METALLIC          2000014000091000000000000000006(2119
ORGANO         METALLOIDAL       2000014000006000000000000000006(32192
ORGANOSILICOCOMPOUNDS            2000007000153000000000000000005(2100
ORIENTATION CONTROL              2000038000703000000000000000002(
ORIENTATION DIFFERENCE           2000038000098000000000000000002(
ORIENTATION RELATIONSHIP         2000038000167000000000000000002(
ORIENTATION  AND                 2000038016622000000000000000004(
ORIENTATION  OF                  2000038030170000000000000000010(
ORIENTATION  ON                  2000038004634000000000000000002(

ORIENTED       MAGNETIC          2000010000298000000000000000002(
ORIENT          THE              2000003008911000000000000000007(
ORIFICE        AREA              2000046000215000000000000000002(
ORIFICE        M                 2000046007364000000000000000002(
ORIFICE        PLATE             2000046000409000000000000000002(
ORIFICE         AND              2000046016622000000000000000002(
ORIFICE         IN               2000046008879000000000000000002(

ORIFICE        SIZE              2000046000207000000000000000002(
ORIFICE        TYPE              2000046000477000000000000000003(61
ORIFICES        AND              2000021016622000000000000000003(
```

Table 1 (Continued)

ORIFICES	FOR	2000021007406
ORIFICES	OF	2000021030170
ORIGINALLY	DEVELOPED	2000006000404
ORIGINALLY	EMPLOYED	2000006000075
ORIGINAL	GERMAN	2000032000072 +585
ORIGINAL	THICKNESS	2000032000225
ORIGINATING	IN	2000004008879
ORIGIN	OF	2000017030170
ORR	SOMMERFELD	2000003000066
ORTHOGONAL	CUTTING	2000019000094 +755
ORTHOGONAL	FUNCTIONS	2000019000175
ORTHOGONAL	METAL	2000019000594
ORTHO	ESTERS	2000011000012
ORTHO	PARAHYDROGEN	2000011000007
ORTHO	PARA	2000011000010 +1653
ORTHOTROPIC	PLATE	2000022000409 +150
ORTHOTROPIC	SHELL	2000022000132
O	ALUMINUM	2000084000470
O	CB	2000084000027 +1190
O	F	2000084000402 +39
O	IRON	2000084000181
O	MN	2000084000038 +422
O	MO	2000084000123
O	RINGS	2000084000038 +422
O	AND	2000084016622
O	AT	2000084002770
O	OR	2000084000906
O	THE	2000084008911
O	SUB3	2000084000006
O	TI	2000084000236 +68
OSCILLATING	AEROFOILS	2000049000018
OSCILLATING	BODY	2000049000301
OSCILLATING	CASCADE	2000049000128
OSCILLATING	LOW	2000049000583
OSCILLATING	BETWEEN	2000049000676
OSCILLATING	IN	2000049008879
OSCILLATING	WITH	2000049003844
OSCILLATING	SORT	2000049000033
OSCILLATING	SURFACE	2000049000641
OSCILLATION	ANALYSIS	2000027001094
OSCILLATION	GIVEN	2000027000494
OSCILLATION	OF	2000027030170
OSCILLATION	WHICH	2000027001180
OSCILLATION	TYPE	2000027000477
OSCILLATIONS+		2000066000426
OSCILLATIONSM		2000066007364
OSCILLATIONS	OBSERVED	2000066000101
OSCILLATIONS	AND	2000066016622
OSCILLATIONS	ARE	2000066000996
OSCILLATIONS	IN	2000066008879
OSCILLATIONS	OF	2000066030170

Table 1 (Continued)

Table 1 (Continued)

OVERALL	SOUND	2000047000125
OVERALL	SYSTEM	2000047000916
OVERALL	TURBINE	2000047000603
OVERCOME	BY	2000019002759
OVEREXPANDED	NOZZLES	2000006000157
OVERHEAT	TEMPERATURES	2000002000726
OVERSIZED	HEADS	2000002000007
OVERSPEED	PROTECTION	2000008000088
OVERTAKING	OF	2000002030170
OVERTEMPERATE	XPOSURES	2000012000013
OVERTEMPERAT	PARTIALLY	2000012000026
OVERVOLTAGE	ON	2000003004634
OWING	TO	2000005008045
OXIDANT	FUEL	2000012000578
OXIDANT	MIXTURE	2000012000088
OXIDATION	BEHAVIOR	2000252000332
OXIDATION	CORROSION	2000252000312
OXIDATION	DATA	2000252000943
OXIDATION	KINETICS	2000252000049
OXIDATION	MAJOR	2000252000089
OXIDATION	MECHANISM	2000252000191
OXIDATION	M	2000252007364
OXIDATION	PROTECTION	2000252000088
OXIDATION	PROTECTIVE	2000252000059
OXIDATION	RATE	2000252000442
OXIDATION	RATES	2000252000208
OXIDATION	RESISTANCE	2000252000356
OXIDATION	RESISTANT	2000252000164
OXIDATION	AND	2000252016622
OXIDATION	AT	2000252002770
OXIDATION	IN	2000252008879
OXIDATION	IT	2000252000260
OXIDATION	OF	2000252030170
OXIDATION	TO	2000252008045
OXIDATION	STABILITY	2000252000414
OXIDATION	STUDIES	2000252000458
OXIDATION	TESTING	2000252000432
OXIDATION	TEST	2000252000775
OXIDATION	TESTS	2000252000835
OXIDATIVE	STABILITY	2000004000414
OXIDE	BODIES	2000185000217
OXIDE	CERAMIC	2000185000169
OXIDE	COATED	2000185000050
OXIDE	COATINGS	2000185000275
OXIDE	CONTENT	2000185000092
OXIDE	FILM	2000185000153
OXIDE	FILMS	2000185000073
OXIDE	GLASSY	2000185000008
OXIDE	INCLUSIONS	2000185000015
OXIDE	LAYER	2000185000634
OXIDE	LOSS	2000185000110
OXIDE	METAL	2000185000594

Table 1 (Continued)

OXIDE	M	2000185007364000000000000000005(
OXIDE	NIOBIUM	2000185000750000000000000000002(
OXIDE	NUCLEATION	2000185000130000000000000000002(
OXIDE	PENETRATION	2000185000330000000000000000002(
OXIDE	AND	2000185016622000000000000000018(
OXIDE	BY	2000185002759000000000000000002(
OXIDE	FROM	2000185001807000000000000000002(
OXIDE	IN	2000185008879000000000000000009(
OXIDE	UNDER	2000185000596000000000000000002(
OXIDE	WITH	2000185003844000000000000000002(
OXIDE	SINGLE	2000185000240000000000000000005(50
OXIDE	SURFACES	2000185000206000000000000000002(
OXIDE	SYSTEM	2000185000916000000000000000003(
OXIDE	SYSTEMS	2000185000757000000000000000002(
OXIDE	TOOLS	2000185000390000000000000000002(
OXIDE	VAPOR	2000185000137000000000000000002(
OXIDES	M	2000057007364000000000000000003(
OXIDES	AND	2000057016622000000000000000008(
OXIDES	AT	2000057002770000000000000000003(
OXIDES	IN	2000057008879000000000000000002(
OXIDES	OF	2000057030170000000000000000006(
OXIDES	ON	2000057004634000000000000000002(
OXIDES	OR	2000057000906000000000000000002(
OXIDES	SUCH	2000057000222000000000000000002(
OXIDES	WITH	2000057003844000000000000000003(
OXIDES	STUDIED	2000057000231000000000000000002(
OXIDIC	REFRACTORY	2000003000194000000000000000003(2319
OXIDIZED	COPPER	2000018000122000000000000000003(614
OXIDIZED	TO	2000018008045000000000000000003(
OXIDIZER	WHICH	2000017001180000000000000000003(67
OXIDIZING	ACIDS	2000008000013000000000000000002(
OXIDIZING	ATMOSPHERES	2000008000046000000000000000002(
OXYACETYLENETORCH		2000003000019000000000000000002(
OXYGEN	AIR	2000236000846000000000000000002(
OXYGEN	ALLOYS	2000236001098000000000000000002(
OXYGEN	CARBON	2000236000214000000000000000002(
OXYGEN	CONCENTRATIO	2000236000145000000000000000007(92
OXYGEN	CONTAMINATIO	2000236000045000000000000000003(129
OXYGEN	CONTENT	2000236000092000000000000000004(82
OXYGEN	DISSOLUTION	2000236000020000000000000000002(
OXYGEN	FLAME	2000236000246000000000000000003(23
OXYGEN	FLUORINE	2000236000050000000000000000004(152
OXYGEN	FREE	2000236000416000000000000000002(
OXYGEN	HYDROGEN	2000236000410000000000000000006(27
OXYGEN	MIXTURES	2000236000117000000000000000012(195
OXYGEN	M	2000236007364000000000000000006(
OXYGEN	NITROGEN	2000236000133000000000000000013(186
OXYGEN	PRESSURE	2000236001366000000000000000004(
OXYGEN	ROCKET	2000236000624000000000000000003(
OXYGEN	AND	2000236016622000000000000000033(
OXYGEN	AS	2000236001689000000000000000003(
OXYGEN	FROM	2000236001807000000000000000002(

Table 1 (Continued)

OXYGEN	IN	20002360088790000000000000000024(
OXYGEN	IS	20002360017040000000000000000002(
OXYGEN	ON	20002360046340000000000000000005(
OXYGEN	TO	20002360080450000000000000000003(
OXYGEN	UNDER	20002360005960000000000000000002(
OXYGEN	WITH	20002360038440000000000000000004(
OXY	ACETYLENE	20000070000110000000000000000003(-17532
OXY	HYDROGEN	20000070004100000000000000000002(
OZONE	FLAMES	20000120000970000000000000000002(
OZONE	OXYGEN	20000120002360000000000000000003(-476
OZONE	AND	20000120166220000000000000000003(

Section IV: CACL-31

NASA VOCABULARY TWO-WORD STRINGS, THEIR USAGE, AND RELATION TO SYSTEM CBU'S IN THE GE-2 AUTO-INDEXED MESSAGE COLLECTION *

A. Introduction

This section reports a study of the two-word strings in the NASA indexing vocabulary and their relationship to the GE-2 prototype retrieval system. The principal motivation and the discussion of the results of the data gathered in this study appear in Chapter VI of the Evaluation Study Report. We wish to note that V. E. Giuliano was a major participant in conducting the study whose details are recorded here. Briefly, the two-word strings in the NASA term list are regarded, in the report, as representatives of Subject Heading queries that might be posed to the GE-2 retrieval system. Their overlap properties with the GE-2A machine-indexing vocabulary is thus of considerable interest. Similarly, it is of importance to determine how many of these two-word NASA terms turn out to be System CBU's in the GE-2 system. The data for determining these relationships are presented here together with other observations about the set of two-word terms. Because of their importance to studies now under way, Appendix A contains an exhaustive listing of the pairs that were studied, together with the observations made on each.

B. NASA Postings to Two-Word Terms

The first step was to estimate the total number of postings of two-word strings in the total NASA collection. Consequently, the NASA frequency-ordered indexing vocabulary list was divided rather arbitrarily into ten intervals. Samples were taken from each of the intervals and were used to estimate the average number of postings to the terms in each frequency interval. From Figure 3 of Technical Note CACL-29, the

* Issued on June 15, 1966 to a limited distribution as Technical Note CACL-31 by Peter R. Bono and Paul E. Jones, Jr. References only have been updated.

average number of two-word strings per page could be calculated (making appropriate allowance for the fact that not all MWT's are two-word strings).

Table A records all the information necessary for the estimation of the total number of postings to two-word strings in each interval. The number of postings is calculated by taking the product of the average number of two-word strings per page and the number of pages in the interval and the average term frequency over the whole interval. For example, for sample 2, we multiply 39 x 6 x 150 and get 35,200 -- the estimated number of postings in the interval of pages 13-18. From Table A, we see that the total estimated number of postings to two-word strings in the NASA collection is 213,505 tokens.

C. Relating NASA Subject Headings to GE Collection Parameters

The next step was to classify the two-word strings into six groups we wished to differentiate. In order to describe these groups, it is helpful to use the following notation. Represent a word pair as $\widehat{ab}$ where "a" and "b" respectively represent the first and second words of the pair. Let f_{ab} denote the frequency in the GE-2 corpus of the two-word string $\widehat{ab}$, and let C_{ab} denote the coherence measure of $\widehat{ab}$
Then the six groups of two-word strings can be described with precision:

TABLE A

NUMBER OF POSTINGS TO MULTIPLE-WORD TERMS

Sample Number	Interval in NASA Vocabulary Book (by page number)	Number of Pages in Interval	Average Number of Two-Word Terms Per Page	Average Term Frequency Over Whole Interval	Estimated Number of Pastings in Interval	Number in Sample
1	p.1-p.12	12	---*	---*	49,705* (actual count)	122
2	p.13-p.18	6	≈39	≈150	35,200	36
3	p.19-p.24	6	≈47	≈ 93	26,200	60
4	p.25-p.36	12	≈54	≈ 57	37,000	72
5	p.37-p.48	12	≈56	≈34	22,800	48
6	p.49-p.72	24	≈59	≈16	22,400	24
7	p.73-p.96	24	≈59	≈ 8	11,200	12
8	p.97-p.120	24	≈59	≈4	5,600	12
9	p.121-p.144	24	≈59	≈1.5	2,100	12
10	p.145-p.169	25	≈52	≈1	1,300	12

Total Estimated Number of Postings = 213,505

*NOTE: The number of posting for the 122 two-word terms in the first twelve pages of the NASA index term dictionary was actually counted; consequently, the two statistics noted above are unnecessary.

Section IV: CACL-31

Group	Definition
G	Word "a" in GE-2A Vocabulary List; word "b" in GE-2A Vocabulary List; $f_{ab} \geq 2$ $C_{ab} \geq 20$
N/A	Word "a" <u>not</u> in GE-2A Vocabulary List; word "b" in GE-2A Vocabulary List
A/N	Word "a" in GE-2A Vocabulary List; word "b" <u>not</u> in GE-2A Vocabulary List
N/N	Word "a" <u>not</u> in GE-2A Vocabulary List; word "b" <u>not</u> in GE-2A Vocabulary List
C	Word "a" in GE-2A Vocabulary List; word "b" in GE-2A Vocabulary List $f_{ab} \geq 2$, but $C_{ab} < 20$
F	Word "a" in GE-2A Vocabulary List; word "b" in GE-2A Vocabulary List <u>but</u> $f_{ab} \leq 1$

This classification of the NASA two-word string permits us to explore the relationship between NASA subject headings and the GE-2 System CBU's. The number of postings NASA had given each string was also of interest in developing this relationship. Accordingly, we proceeded to estimate the total NASA usages (postings) for each of the six kinds of strings. Using the same data as used in the preparation of Table A, we calculated this figure by taking the number of two-word strings of each kind in the sample for our interval, dividing by the sample size for the interval and multiplying by the estimated total number of

postings to two-word strings in that interval. For example, in sample 2 (pp.13-18), 10 of the 36 sample strings belong to group G. The total estimated number of postings for this interval was 35,200 (cf. row 2, Table A). Consequently, the total estimated number of postings for group G for this interval is

$$10 \times \frac{1}{36} \times 35{,}200 = 9{,}950$$

Table B-I displays these data for each of the six groups and for each of the ten intervals. Also, the ratio, P_k, of total number in Group X to the total size of sample k is recorded (where X = G, A/N, N/A, C, or F; and k = 1, . . ., 10). In the above example, this would be

$$P_2 = \frac{10}{36} = .28$$

D. Measures Reflecting the NASA/GE Relationship

We also wished to have an estimate of the number of types (as opposed to tokens) for each group of two-word strings. The total number of types for each page interval was estimated using Figure 3 of TN CACL-29 (again multiplying the total number of MWT's by an appropriate scaling factor to allow for the fact that some of the MWT's are composed of more than two words).

The estimates of the number of types for each group were calculated simply by multiplying the total estimated number of types by the ratio P_k (from Table B-I). These figures are displayed in Table B-II. For example, in sample 2, there are approximately 234 two-word strings. Since 10 of the 36 strings sampled belonged to Group G, we would expect that $\frac{10}{36} \times 234 = P_2 \times 234 = 65$ of the 234 would belong to Group G.

TABLE B-I

Sample Number	Total Est. Number of Postings	Total Sample Size		G	N/A	A/N	N/N	C	F
1	49,705 (actual)	122	No. in Sample	67	27	7	4	8	9
			Actual No. of Postings	33,691	8,423	2,102	1,303	2,206	1,980
			$P_1 = \frac{\text{Act. No. of Postings}}{\text{Total No. of Postings}}$	.68	.12	.06	.03	.06	.05
2	35,200	36	No. in Sample	10	9	6	1	7	3
			Est. No. of Postings	9,900	8,800	6,000	1,000	6,700	2,800
			$P_2 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.28	.25	.17	.03	.19	.08
3	26,200	60	No. in Sample	22	6	11	4	9	8
			Est. No. of Postings	9,700	2,620	4,700	1,830	3,950	3,400
			$P_3 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.37	.10	.18	.07	.15	.13
4	37,000	72	No. in Sample	19	18	9	10	8	8
			Est. No. of Postings	10,000	9,250	4,450	5,300	4,000	4,000
			$P_4 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.27	.25	.12	.14	.11	.11
5	22,800	48	No. in Sample	9	15	9	4	4	7
			Est. No. of Posting	4,300	7,100	4,300	1,900	1,900	3,300
			$P_5 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.19	.31	.19	.083	.083	.144

TABLE B-I (continued)

6	22,400	24	No. in Sample	3	7	3	2	2	7
			Est. No. of Postings	2,800	6,500	2,800	1,900	1,900	6,500
			$P_6 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.125	.29	.125	.083	.083	.29
7	11,200	12	No. in Sample	1	6	1	1	0	3
			Est. No. of Postings	933	5,600	933	933	0	2,800
			$P_7 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.083	.50	.083	.083	.00	.25
8	5,600	12	No. in Sample	0	6	3	1	0	2
			Est. No. of Postings	0	2,800	1,400	470	0	930
			$P_8 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.00	.50	.25	.083	.00	.167
9	2,100	12	No. in Sample	0	6	3	3	0	0
			Est. No. of Postings	0	1,050	525	525	0	0
			$P_9 = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.00	.50	.25	.25	.00	.00
10	1,300	12	No. in Sample	0	8	1	1	0	2
			Est. No. of Postings	0	864	109	109	0	218
			$P_{10} = \frac{\text{No. in Sample}}{\text{Total Sample Size}}$	.00	.667	.083	.083	.00	.167
	213,505	400	Total of Est. Postings	71,324	53,007	27,319	15,270	20,656	25,928

TABLE B-II

Sample Number	Total Est. Number of Postings	Est. Total Number of Types		G	N/A	A/N	N/N	C	F
1	49,705 (actual)	122 (actual)	Actual No. of Types	67	27	7	4	8	9
2	35,200	234	Est. No. of Types	65	59	39	7	46	18
3	26,200	282	Est. No. of Types	104	28	51	20	42	37
4	37,000	658	Est. No. of Types	178	165	79	92	72	72
5	22,800	660	Est. No. of Types	125	203	125	56	56	95
6	22,400	1,400	Est. No. of Types	175	408	175	117	117	408
7	11,200	1,400	Est. No. of Types	116	700	117	117	0	350
8	5,600	1,400	Est. No. of Types	0	700	350	116	0	234
9	2,100	1,400	Est. No. of Types	0	700	350	350	0	0
10	1,300	1,250	Est. No. of Types	0	833	104	104	0	209
	213,505	8,806	Column Totals	830	3,823	1,397	983	341	1,432

Section IV: CACL-31

There are a number of significant measures which can be calculated from data in Tables B-I and B-II. These measures -- four in all -- are defined as follows:

$\alpha = \frac{G + C + F}{N}$ = Of the total number of two-word strings, the percentages of strings $\widehat{ab}$ such that both "a" and "b" are members of the GE-2A Machine Indexing Vocabulary

$\beta = \frac{G + C}{G + C + F}$ = Of the total number of two-word strings $\widehat{ab}$ such that "a" and "b" are members of the GE-2A vocabulary, the percentage of strings such that $f_{ab} \geq 2$

$\gamma = \frac{G}{G + C + F}$ = Of the total number of two-word strings $\widehat{ab}$ such that "a" and "b" are members of the GE-2A vocabulary, the percentage of strings such that $f_{ab} \geq 2$ and $C_{ab} \geq 20$

$\delta = \frac{G + C + F + A/N + N/A}{N}$ = Of the total number of two-word strings the percentage of strings such that at least one of the words is in GE-2A.

[Note: N = total number of tokens (or types)]

Section IV: CACL-31

These measures have been calculated for both tokens and types and are displayed in Table B-III.

Finally, the percentages (P_k) given in Table B-I were plotted versus the term usage frequency in the NASA collection (equivalent to the page number in the printout of the frequency-order NASA 18K index term dictionary).

The graph was produced by a smoothing procedure which took an average of the P_k's over three consecutive values. That is, the new

$$P_k = P_k' = \frac{P_{k-1} + P_k + P_{k+1}}{3}$$ for k=1, --, 9. The final point is smoothed by taking

$$P_{10}' = \frac{P_9' + P_{10}}{2}$$

These smoothing calculations are recorded in Table C-I.

From the resulting graph (Table C-II), it can be seen that for two-word strings that have a high frequency of usage, the probability is quite high that both words of the string will belong to the GE-2A Vocabulary and that its $f_{ab} \geq 3$ and $C_{ab} \geq 20$ (i.e. the string belongs to group G). However, as we consider strings with lower and lower NASA frequency, this probability falls practically to zero. That is, group G draws most of its members from the high-frequency two-word strings. The graph also shows that, except for initial disturbances among the high frequency strings, groups A/N and N/N draw equally from the whole NASA indexing vocabulary. That is, these two groups appear to be largely independent of NASA frequency of usage.

TABLE B-III

α-, β-, γ-, AND δ-TYPE MEASURES CALCULATED FOR THE NASA INDEX VOCABULARY FOR BOTH TOKENS AND TYPES

	TOKENS	TYPES
α	$\frac{117,908}{213,505} \approx .552$	$\frac{2,603}{8,806} \approx .296$
β	$\frac{91,980}{117,908} \approx .780$	$\frac{1,171}{2,603} \approx .450$
γ	$\frac{71,324}{117,908} \approx .605$	$\frac{830}{2,603} \approx .319$
δ	$\frac{198,235}{213,505} \approx .933$	$\frac{7,823}{8,806} \approx .888$

TABLE C-I

SMOOTHED DATA FOR THE SIX CURVES PRESENTED IN TABLE C-II

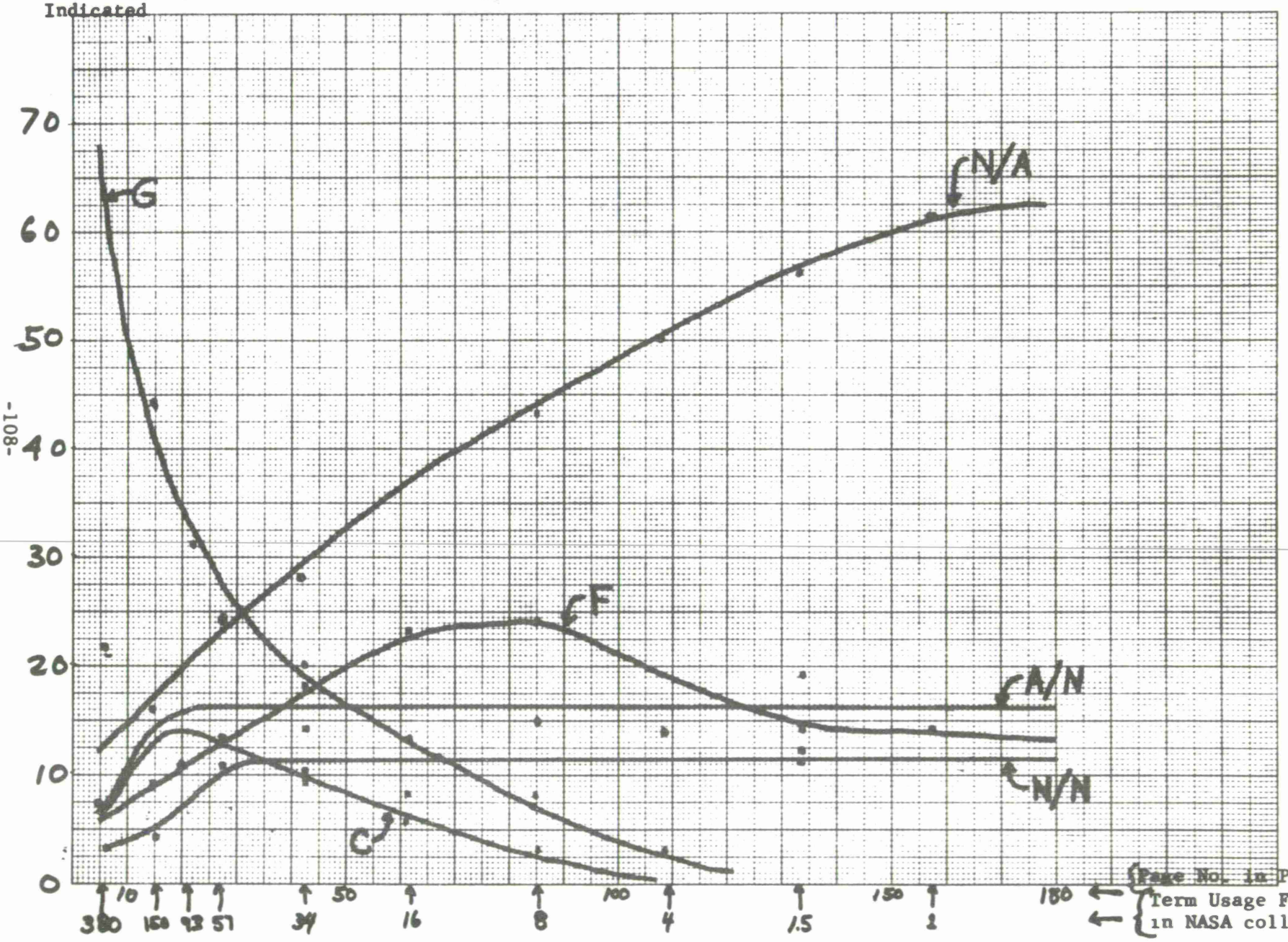
	G		N/A		A/N		N/N		C		F	
	P_k	P'_k	P_k	P'_k	P_k	P'_k	P_k	P'_k	P_k	P'_k	P_k	P'_k
1	.68	.68	.12	.12	.06	.06	.03	.03	.06	.06	.05	.05
2	.28	.44	.25	.16	.17	.14	.03	.04	.19	.13	.08	.09
3	.37	.31	.10	.20	.18	.16	.07	.08	.15	.15	.13	.11
4	.27	.24	.25	.23	.12	.16	.14	.10	.11	.11	.11	.13
5	.19	.20	.31	.28	.19	.14	.083	.10	.083	.09	.14	.18
6	.13	.13	.29	.37	.12	.13	.083	.08	.083	.06	.29	.23
7	.083	.07	.50	.43	.083	.15	.083	.08	.00	.03	.25	.24
8	.00	.03	.50	.50	.25	.19	.083	.14	.00	.00	.17	.14
9	.00	.00	.50	.56	.25	.19	.25	.14	.00	.00	.00	.11
10	.00	.00	.67	.61	.083	.14	.083	.11	.00	.00	.17	.14

for k=1, --, 9; $P'_k = \frac{P_{k-1} + P_k + P_{k+1}}{3}$, where P_k = observed percentage at "time" k

k=10; $P'_{10} = \frac{P'_9 + P_{10}}{2}$

TABLE C-II

% of Star Two-Word Subject Heading in Page (or Frequency) Interval Indicated of Type Indicated



APPENDIX A

SAMPLE 1

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
2264	heat transfer	630	G	1547	791	231
2218	high temperature		G			57
1462	rocket engine		G			100
1328	space flight	20	G	450	374	54
967	space vehicle	71	G	450	259	280
932	solid propellant	105	G	431	384	290
877	high energy		G			34
843	radiation effect		G			24
778	magnetic field		G			464
764	cross section		G			1047
706	mach number		G			524
679	wave propagation	23	G	371	174	160
600	computer program		G			232
586	high speed		G			88
586	high frequency		G			33
564	cosmic radiation		N/A			626
558	reynolds number		G			541
549	space environment	11	G	450	115	96
544	manned spacecraft		N/N			
504	low temperature		G			54
504	upper atmosphere		N/A			
489	differential equation		G			328
475	high altitude		G			38
467	carbon dioxide		A/N			1248
450	launch vehicle		N/A			1008
447	reentry vehicle		N/A			
440	hypersonic flow		G			59
436	gas flow		G			22
427	data processing		G			38
426	supersonic flow	119	G	517	2402	43
422	aluminum alloy		G			133
419	electronic equipment		G			243
418	VTOL aircraft	32	G	94	636	230
416	apollo project		N/A			
415	doppler effect		N/A			
410	automatic control		G			131
406	aerodynamic characteristics		G			81
404	refractory metal		G			278
398	geomagnetic field		N/A			
398	wind tunnel	274	G	303	407	1000
396	electromagnetic wave		G			138
391	low frequency		G			73
382	gamma radiation		N/A			

Appendix A

SAMPLE 1 (Continued)

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
380	turbulent flow	56	G	316	2402	33
377	thin film	9	G	221	226	81
369	rocket nozzle		G			110
368	boundary layer		G			709
367	flight test		G			43
367	optimal control		N/A			
366	NASA program		N/A			
359	supersonic transport	19	G	517	172	96
350	aircraft design	14	C	636	1173	
350	high pressure	72	C	1757	1563	
349	blunt body		G			436
347	aerospace medicine		N/N			
346	mechanical property		G			264
338	solar flare		A/N			
337	single crystal	43	G	24	105	7670
331	electron beam		G			554
331	material testing	3	C	102	432	
327	communications satellite		N/A			
318	plastic deformation		G			351
317	materials science		A/N			
312	shock tunnel	16	G	597	336	36
311	liquid propellants		G			100
308	charged particle		N/A			
307	high strength		G			54
303	satellite observation		F			
303	satellite orbit	3	G	77	21	835
300	attitude control		N/A			
300	laser output		N/A			
298	low pressure		G			23
297	digital computer		G			1720
293	gas dynamics		G			97
293	titanium alloy	146	G	599	1691	65
289	fluid mechanics		A/N			
286	electron density		G			76
283	solar radiation	8	G	88	303	133
283	space science		A/N			
282	temperature effect	12	C	2033	1866	1.4
280	cylindrical shell		G			413
278	human performance		N/A			
278	physiological response		N/A			
278	radiation measurement		F			
276	temperature measurement	32	C	2033	704	10
275	mercury project		N/A			
275	spacecraft propulsion		N/A			
274	infrared radiation		N/A			
274	reliability engineering		F			

Appendix A

SAMPLE 1 (Continued)

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
272	motion equation		F			
271	communication system		N/A			
270	energy conversion		G			167
218	temperature distribution	93	G	2033	550	38
217	crack propagation		G			984
215	E layer		F			
215	planetary atmosphere		N/A			
213	boundary value		G			113
213	flight control	2	C	374	37	
213	liquid metal		G			64
213	shell theory	3	C	132	808	13
212	chemical kinetics		G			315
211	Maxwell equation		N/A			
210	radioactive isotope		N/N			
210	satellite measurement		F			
209	computer method		F			
209	linear system		G			23
209	thermal stress	55	G	633	1227	32
208	tensile strength	68	G	359	761	112
207	control device		G			
207	satellite communication		A/N			
206	radar system		N/A			
203	aerodynamic heating		G			290
202	human tolerance		N/N			
202	ultraviolet radiation		N/A			
201	nuclear explosion		A/N			
201	transport aircraft	13	G	151	636	59
200	niobium alloy	3	C	75	1098	
200	orbit calculation		F			
199	atmospheric composition		F			
199	environmental testing		N/A			
198	meteorological satellite		N/A			
197	gravitational field		N/A			

Appendix A

SAMPLE 2

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
196	elastic defomation		G			20
195	high power		C			
194	absorption spectrum		A/N			454
194	glass fiber		G			866
194	nozzle flow		C			
194	power generator		C			
178	gas discharge		G			
177	ionizing radiation		N/A			626
177	low density		G			124
177	radio wave		N/A			
177	satellite tracking		A/N			
176	test facility		G			
161	human body		N/A			
160	acceleration stress		F			
160	nuclear propulsion		G			81
160	pressure measurement		C			
160	time dependency		A/N			
159	radiation shielding		A/N			
146	polymer chemistry		A/N			
145	dynamic stability		G			34
145	integral equation		G			199
145	Newton Theory		N/A			
145	STOL aircraft		N/A			
144	electric discharge		N/A			
134	rare earth		N/A			3985
133	gas mixture		G			63
133	radiation intensity		F			
133	shell stability		C			
132	analog computer		G			1085
132	compressible flow		C			
124	aircraft performance		C			
124	Lagrange equation		N/A			343
124	mass spectrometry		A/N			
124	molecular structure		F			
124	transonic speed		G			
124	celestial mechanics		N/N			

Appendix A

SAMPLE 3

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
115	heat exchanger		G			250
115	Liapunov function		N/A			
115	temperature control	10	C	2033	703	3.5
115	transport property	29	G	151	1350	64
114	atmospheric density		F			
107	Defender project		N/A			
107	magnesium oxide		G			164
107	matrix analysis		F			
107	optical pumping		A/N			
107	phase shift		A/N			
100	radiation field		G			54
100	solar system		F			
99	human engineering		A/N			
99	ionospheric sounding		N/N			
99	metal surface		G			51
94	structural beam		F			
93	atmospheric temperature		F			
93	axisymmetric flow		C			
93	checkout equipment		N/A			
93	chromium alloy		G			26
89	orbital element		N/A			
89	pattern recognition		A/N			
89	satellite control		F			
89	structural engineering		F			
88	atmospheric ionization		A/N			
83	mercury capsule		N/N			
83	solar spectrum		A/N			
83	solar eclipse		A/N			
83	wave interaction	3	G	371	119	31
82	ground station		A/N			
112	wave diffraction	3	G	371	61	60
111	flow characteristics		C			
111	ionospheric storm		N/N			
111	jet engine		G			101
111	nimbus satellite		N/A			
104	static stability	2	G	23	414	95
104	vortex flow	6	C	106	2184	12
103	arc jet		G			46
103	approximation method		G			15
103	heat flux		G			254
97	molecular beam		G			243
97	optical property		F			
97	propellant combustion		C			
97	radiation transfer		C			
97	spherical shell	21	G	88	255	421

Appendix A

SAMPLE 3 (Continued)

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
92	thrust chamber	12	G	418	266	46
91	helicopter rotor		G			1236
91	high performance		C			
91	ion source		G			152
91	jet aircraft		C			
86	oxidation resistance		G			125
86	power plant		G			565
86	space biology		A/N			
86	test method	19	C	775	2040	
86	transmission line	4	G	61	157	188
81	galactic radiation		N/A			
81	Lorentz transformation		N/A			1397
81	lunar spacecraft		N/N			
81	radiation medicine		A/N			
81	solar proton		A/N			

SAMPLE 4

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
79	velocity profile	38	G	541	161	195
79	velocity measurement	4	C	541	437	7.6
79	viscous fluid	14	G	138	455	100
74	biological cell		N/A			
74	catalytic activity		N/N			
74	elastic shell		G			163
70	power source		G			35
70	probability distribution		N/A			
70	transient response	20	G	157	202	284
67	satellite perturbation		F			
67	steady flow	34	G	229	2402	28
67	surface reaction	5	G	641	113	31
63	orbital launch		N/N			
63	orbital motion		N/A			
63	plastic flow		C			
60	static testing	2	G	23	432	91
60	vibration effect		F			
60	wave attenuation		A/N			
57	Michigan project		N/A			
57	military technology		N/N			
57	plasma arc		G			46

Appendix A

SAMPLE 4 (Continued)

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
54	aircraft noise		G			48
54	beryllium hydride		A/N			
54	electrochemical cell		N/A			
52	hypervelocity projectile		N/N			
52	indium antimonide		N/N			
52	information retrieval		A/N			
50	radioactive material		N/A			
50	random vibration		G			273
50	remote control		N/A			
48	state equation		F			
48	structural reliability		F			
48	thermal convection		F			
46	reactor safety		G			280
46	reentry condition		N/A			
46	reinforcing fiber		N/A			
76	cold working		G			419
76	corrosion prevention		A/N			
76	Euler equation		N/A			914
72	microwave radiation		N/A			
72	neutron scattering		N/N			
72	phased array		N/N			
68	electromagnetic interaction		F			
68	flow pattern		C			
68	gaseous laser		A/N			
65	optical method		C			
65	radar measurement		N/A			
65	scatter propagation		N/A			
62	thermal expansion	25	G	633	145	123
61	cardiovascular system		N/A			
61	coriolis effect		N/A			
58	atmospheric moisture		A/N			
58	bending moment		G			518
58	combustion product		G			146
56	missile control		C			
56	plasma confinement		A/N			
56	pressure transducer		A/N			69
53	Debye temperature		N/A			
53	earth crust		A/N			
53	ion density		C			
51	radioactive fallout		N/N			
51	radiation spectrum		A/N			
51	spacecraft stability		N/A			
49	plasma engine		C			
49	simulated altitude	6	G	57	178	256
49	synoptic meteorology		N/N			
47	rotor aerodynamics		F			
47	rotating fluid		C			
47	signal distortion		N/N			
41	control system		G			54
41	delay line		N/A			
41	dynamic model		F			

Appendix A

SAMPLE 5

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
44	proton energy		N/A			
44	seasonal variation		N/A			
42	elliptical orbit		N/A			
42	energy exchange		A/N			
40	elastic bending		F			
40	elementary particle		N/A			
38	aerospace system		N/A			
38	air inlet		G			85
37	phase transformation		G			67
37	piezoelectric crystal		N/A			
36	Voyager project		N/A			
35	aircraft production		F			
34	lift fan		G			204
34	meteor shower		N/N			
33	Riemann integral		N/A			
33	scientific satellite		N/A			
32	trace contaminant		N/N			
32	transpiration cooling		N/A			
30	conducting media		F			
30	cyclotron radiation		N/A			
29	data correlation		F			
29	dynamic pressure		G			31
28	hydrogen fluoride		A/N			
28	hypersonic nozzle		C			
43	notch strength		G			21
43	parabolic equation		N/A			
41	control system		G			54
41	delay line		N/A			
39	atmospheric electricity		A/N			
39	beryllium compound		C			
38	storage battery		A/N			
38	submillimeter wave		N/A			
36	fluorescent emission		N/N			
36	flight training		A/N			
35	particle emission		A/N			
35	periodic oscillation		N/A			
33	air cooling		C			
33	boron nitride		G			1855
32	high gain		A/N			
32	hydraulic equipment		F			
31	line spectrum		A/N			
31	linear accelerator		A/N			
30	probability density		G			200
30	radio transmitter		N/N			
29	stress wave	2	C	890	371	
29	surface energy		F			
28	thermal shock	35	G	633	597	
28	titanium oxide		F			

SAMPLE 6

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
27	ground test		C			
26	earth motion		F			
25	broad band amplifier		N/A			
25	thrust measurement	3	C	418	437	714
24	pressure oscillation		C			34
23	logic network		N/N			
22	Cosmos satellite		N/A			
22	sensory deprivation		N/N			
21	ground control		F			
21	toroidal shell		N/A			
20	nickel compound		F			
19	electric potential		F			
19	shear strength	12	G	201	689	39
18	lateral control		N/A			
17	anisotropic shell		N/A			
17	organic coolant		A/N			
16	compression buckling		F			
16	nuclear effect		F			
16	xenon light		N/A			
15	lithium alloy		F			
15	vacuum melting	4	G	176	131	78
14	gravity center		N/A			
14	solar observer		A/N			
13	electron recombination		A/N			

SAMPLE 7

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
13	zirconium compound		F			
12	Oseen approximation		N/A			
11	Feynman diagram		N/A			
11	video equipment		N/A			
10	oxygen recombination		A/N			
9	decision element		N/A			
9	smoke trail		N/N			
8	Haynes alloy		N/A			
8	terrier missile		N/A			
7	fuel pump		G			29
7	rocket project		F			
6	conversion table		F			

Appendix A

SAMPLE 8

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
6	Gulliver program		N/A			
6	plutonium compound		N/A			
6	X band		A/N			
5	hermetical seal		N/A			
5	Rayleigh member		N/A			
4	aircraft antenna		A/N			
4	gas evacuation		F			
4	period equation		F			
4	thrust termination		A/N			
3	celescope project		N/A			
3	Herzberg band		N/A			
3	negative conductance		N/N			

SAMPLE 9

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
3	sodium gallate		A/N			
2	ammonium picrate		N/N			
2	Delilah project		N/A			
2	Hill method		N/A			
2	Multhopp method		N/A			
2	quadranted meteor		N/N			
2	success project		N/A			
1	aircraft accessory		A/N			
1	Cepheus constellation		N/N			
1	Dyson Theory		N/A			
1	Mellas region		N/A			
1	lead acetate		A/N			

SAMPLE 10

NASA f	Term Name	f_{ab}	CBU	f_a	f_b	C_{ab}
1	muscular function		N/A			
1	Piapacs project		N/A			
1	scale error		F			
1	swordfish operation		N/A			
1	Vintis Theory		N/A			
0	Cassiopeia constellation		N/N			
0	Ekman layer		N/A			
0	high volume		F			
0	MAC project		N/A			
0	organic laser		A/N			
0	retargeting missile		N/A			
0	surgical instrument		N/A			

DISTRIBUTION OF C_{ab} VALUES IN A SAMPLE OF CBU MASTER PAIR LIST*

Every 55th entry (the one at the top of the page) in the alphabetical interval A through P in this list was sampled. The entries represent all word pairs ab in GE-2 for which

$$f_{ab} \geq 3$$

$$f_a \leq 2040$$

$$f_b \leq 2040$$

$$C_{ab} = \frac{f_{ab}}{f_a \cdot f_b}, \quad 450{,}000 \geq 20$$

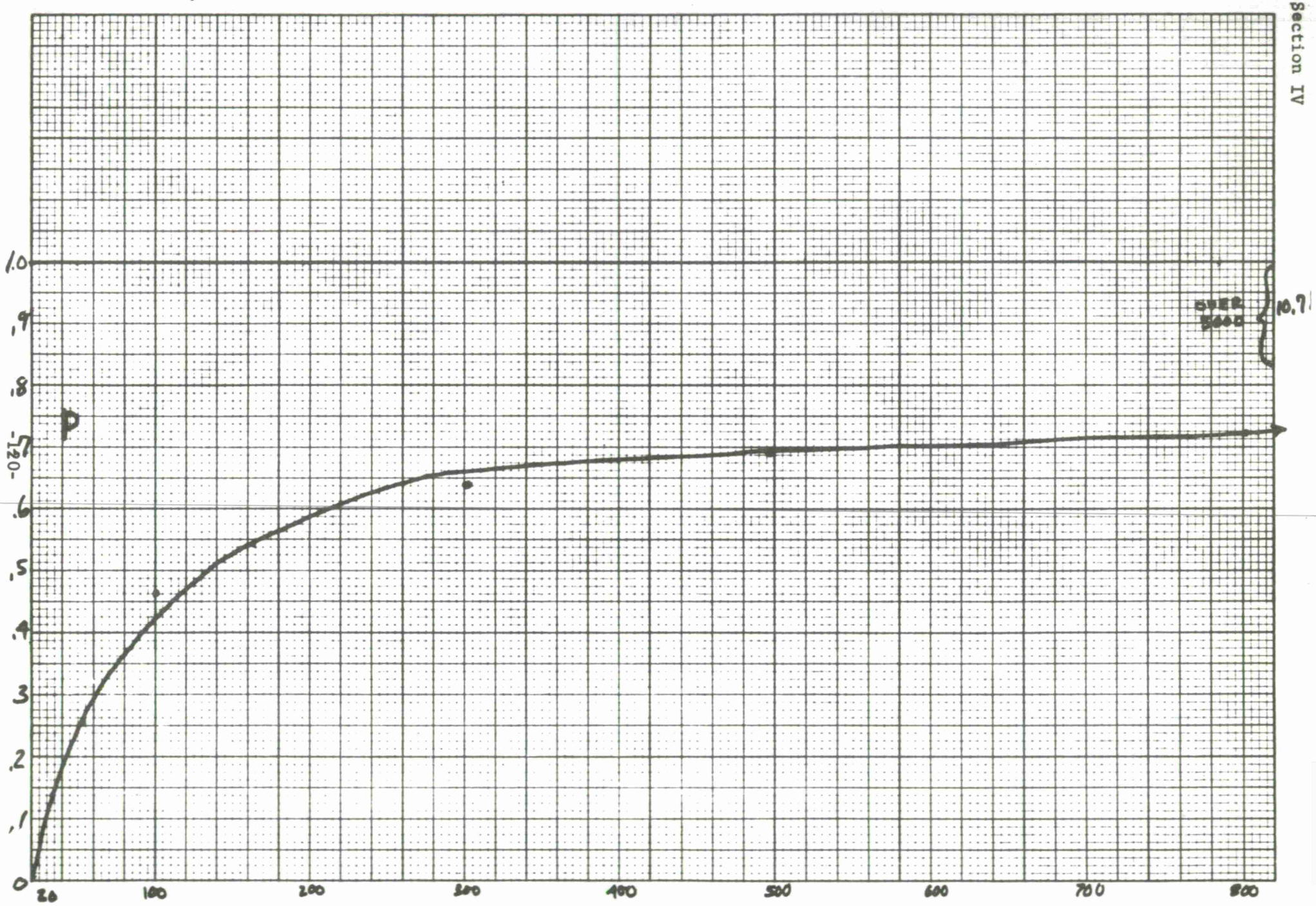
Note that frequencies are not coalesced.

The number of entries with C_{ab} values in certain ranges were tallied. The probability that a pair meeting the above criteria will have in C_{ab} in the stated range was calculated.

C_{ab} Interval	Interval size	No. Entries	$P = \frac{\text{No. Entries}}{94}$
20-49	30	23	.25
50-99	50	20	.21
100-199	100	9	.094
200-299	100	8	.085
300-499	200	5	.053
500-999	500	6	.064
1000-1999	1000	5	.053
2000-4999	3000	8	.085
5000 +	--	10	.107
		94	

* Not previously issued. By Vincent E. Giuliano and Paul E. Jones

Probability that a Word Pair (in the GE 2-A - Master CBU list) will have Cab $\leq$ X



SUMMARY OF DATA PRINTOUTS RETAINED

The list below contains a brief description of the more important data collected in printout form.

1. A listing of all GE-1 documents with the set of GE-1A index terms assigned.
2. A listing of GE-1A index terms with the set of documents to which they were assigned.
3. A listing of the words appearing in the 10,000 abstract GE-2 collection and their frequencies.
4. A listing for each word pair in the GE-2 collection, the frequency of the pair and the frequencies of the two words.
5. A listing of 1, 2, 3, and 4 word strings in frequency order.
6. A listing for each frequency of the number of word types with that frequency.
7. An alphabetic listing of all 3 and 4 word strings appearing 3 or more times with frequencies of constituent substrings given.
8. An alphabetic listing of all word pairs designated as content bearing units.
9. A listing in intervals of 200, of the number of abstracts with a given length in the interval and cumulative, taken in accession number order.
10. A listing of association profiles for each of the 1000 GE 2 terms based on various matrices.

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) Arthur D. Little, Incorporated Cambridge, Mass.	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP N/A

3 REPORT TITLE

PAPERS ON AUTOMATIC LANGUAGE PROCESSING
SELECTED COLLECTION STATISTICS AND DATA ANALYSES

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

5 AUTHOR(S) (Last name, first name, initial)

Jones, Paul E.
Giuliano, Vincent E.
Curtice, Robert M.

6 REPORT DATE February 1967	7a TOTAL NO OF PAGES	7b NO OF REFS
8a CONTRACT OR GRANT NO AF 19(628)-3311 b PROJECT NO c d	9a ORIGINATOR'S REPORT NUMBER(S) ESD-TR-67-202, Vol. I 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None	

10 AVAILABILITY/LIMITATION NOTICES

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, USAF, L. G. Hanscom Field, Bedford, Mass. 01730

13 ABSTRACT

As part of a research program aimed at determining the parameters influencing the effectiveness of a message retrieval system, a collection of 10,000 technical abstracts was indexed and retrieval experiments were conducted with them. Since part of the work involved the development and test operation of an associative retrieval system, basic data about the distribution of words and word strongs were gathered in preparing the system for test and trial. These statistics were thought to be of possible interest to other workers in the field and are gathered as a series of loosely connected papers in this volume under the following groupings: Characteristics and Indexing of GE Data Base; Comparison of Manual and Machine Selected Vocabularies; Vocabulary Distribution Studies; and Studies of Content Bearing Units in Text.

DD FORM 1 JAN 64 1473

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Language English Text Statistics Automatic Indexing Data Processing Word Frequency						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through _____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through _____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through _____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13 ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

GPO 886-551